Psychometrika

CONTENTS

VOLUME THIRTEEN JUNE 1948 NUMBER TWO

The following persons have been nominated for two vacancies in the Council of Directors in the Psychometric Society:

P. J. RULON

G. BENNETT

M. D. ENGELHART

L. R. TUCKER

J. C. FLANAGAN

PSYCHOMETRIKA—VOL. 13, NO. 2
JUNE, 1948

A RATIONALE FOR THE MEASUREMENT OF TRAITS IN INDIVIDUALS*

CLYDE H. COOMBS
UNIVERSITY OF MICHIGAN

Hypotheses presented in a previous paper conceive of an individual's position on an attitude scale as represented by a mean position (status score) and a sigma (dispersion score). The testing of these hypotheses requires a determination of these two scores for each individual. A rationale is presented here for the determination of these two scores for an individual from his forced choice responses to pairs of items of nearly equal scale value.

Certain hypotheses concerning the measurement of psychological traits have been presented in a previous paper.† These hypotheses were based on the two fundamental parameters of an individual obtainable from his performance on a mental test, attitude scale, neurotic inventory, or other suitably designed instruments. The two parameters were called the individual's status score and dispersion score and were each given particular psychological interpretation. Further hypotheses were developed based on the distribution of each of these parameters obtained from a test, scale, or inventory administered to a group of individuals.

The basic problem is the determination of the individual's two parameters in a reliable manner. This is particularly difficult for the dispersion score. The individual's status score is a result of the individual's response to a relatively large number of items, whereas the dispersion score may be reflected in only a small fraction of all the items. The purpose of this paper is to develop a rationale such that the opinions of an attitude scale or the items of a neurotic inventory may be presented to each subject in the form of paired comparisons and both a status and dispersion score may be approximated. Suggestions are also made for a type of mental ability test which would also permit a more reliable determination of dispersion scores.

Another advantage to such a technique lies in the degree to which it removes voluntary control of the individual's score. It is

* This study was made possible by the Bureau of Psychological Services, Institute for Human Adjustment, Horace H. Rackham School of Graduate Studies, University of Michigan.

† Coombs, C. H. Some hypotheses for the analysis of qualitative variables. *Psychol. Rev.*, (in press).

well recognized that an intelligent individual can appear "neurotic" or "stable" as he wills on most existing inventories. Similarly, an individual responding to a set of opinions on an attitude scale can voluntarily choose to have a "pro" or "con" attitude independently of his "true" attitude. However, by presenting an individual with pairs of opinions or statements of the proper distance apart on the scale and forcing a choice between them, he finds himself in a position in which he is less able to guess the "right" answer for any ulterior purpose.

The point from which we begin is the completed attitude scale or inventory with the items' scale positions determined by one of the usual scaling methods. It will be assumed that each item has an exact and known scale position and that its dispersion about that scale position for any cause peculiar to itself (such as ambiguity, etc.) is zero.

Let us consider that the items run the gamut from "pro" to "con." It is postulated further that an individual also has a scale position (S_i) on this same continuum. The individual's scale position, however, is not necessarily a stable one but under different stimulus situations fluctuates about his S_i. We shall assume that the frequency distribution of scale positions taken by an individual follows the normal law (cf. Fig. 1). The mean of this distribution we shall designate as his status score (S_i) and the standard deviation of this distribution as his dispersion score (D_i).

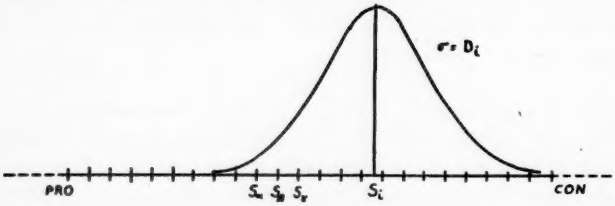


FIGURE 1

Let us designate stimuli with successive scale positions as α, β, γ, δ, $\cdots$, and their respective scale values as S_α, S_β, S_γ, S_δ, $\cdots$. Let us combine these items in pairs such that the members of a pair are not obviously discriminable and administer the inventory with instructions for the individual to select that member of each pair most nearly expressing his own attitude.

We now make the further assumption that an individual chooses that member of a pair of statements which is nearest his own momentary position on the attitude scale.

Theoretically, one could present any pair of stimuli, γ and δ, a large number of times, and from the proportion of time ($P_{\gamma p\delta}$), S_γ was chosen in preference to S_δ, the x/σ value ($Z_{\gamma\delta}$) corresponding to this proportion may be readily obtained from normal probability tables. The following equation may then be written:

$$Z_{\gamma\delta}=\left(\frac{S_\gamma+S_\delta}{2}-S_i\right)\frac{1}{D_i}. \tag{1}$$

Where S_γ and S_δ have been previously determined by scaling, $Z_{\gamma\delta}$ is known, and S_i and D_i are the two unknowns.

Obviously, with a similar equation from another pair of stimuli, λ and μ,

$$Z_{\lambda\mu}=\left(\frac{S_\lambda+S_\mu}{2}-S_i\right)\frac{1}{D_i}, \tag{2}$$

it would be easy to solve for S_i and D_i. By subtracting equation (2) from equation (1) and solving for D_i:

$$D_i=\frac{S_\gamma+S_\delta-(S_\lambda+S_\mu)}{2(Z_{\gamma\delta}-Z_{\lambda\mu})}, \tag{3}$$

and by substituting its value from equation (3) for D_i in equation (1) or (2), S_i would be immediately given.

Application of the above formulas, however, would require that successive judgments on the same pair of stimuli be independent of each other. In the situation in which stimuli are statements of opinion, this condition would not hold. Consequently, some other device must be developed to determine $P_{\gamma p\delta}$.

Let us consider the pair of stimuli γ and δ. It is desired to determine $P_{\gamma p\delta}$ in spite of the fact that γ and δ can be presented to an individual only once. Let us designate all other pairs of stimuli λ and μ such that λ is the member of the pair that has the lower scale value. For purposes of simplicity, we shall regard the origin as being on the left end of the scale and the scale values of the stimuli as increasing to the right.

On the basis of the assumptions made so far, the response of an individual to stimuli λ and μ may be reinterpreted and adjusted in a quantitative manner so that it may be regarded as a response to another pair of stimuli, γ and δ. Suppose, for example, that λ and μ are below γ and δ in scale value and that an individual states that he prefers λ to μ. The scale position of such an individual at that moment is below the abscissa $\dfrac{S_\lambda+S_\mu}{2}$. In order for an individual to make the judgment γ preferred to μ, $(\gamma p \mu)$, he must be below the point $\frac{S_\gamma + S_\delta}{2}$ (cf. Fig. 2). In this instance, then, that $\frac{S_\lambda + S_\mu}{2} < \frac{S_\gamma + S_\delta}{2}$ and $\lambda \; p \; \mu$, this judgment on this pair of stimuli may be regarded as a judgment $\gamma \; p \; \delta$.

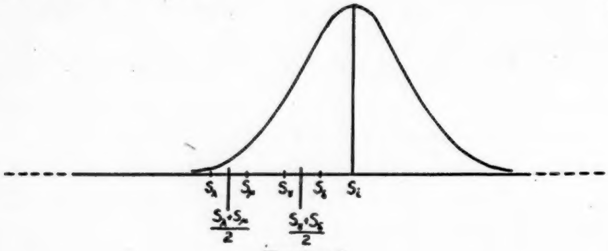

S_λ
S_μ
S_γ
S_δ
S_i
$\frac{S_\lambda + S_\mu}{2}$
$\frac{S_\gamma + S_\delta}{2}$

FIGURE 2

In a similar manner, it may be shown that if $\frac{S_\lambda + S_\mu}{2} > \frac{S_\gamma + S_\delta}{2}$ and $\mu \; p \; \lambda$, then this judgment may be regarded as a judgment $\delta \; p \; \gamma$.

The frequency with which pairs of stimuli λ, μ such that $\frac{S_\lambda + S_\mu}{2} < \frac{S_\gamma + S_\delta}{2}$ and $\lambda \; p \; \mu$ occur will be designated $f_{\lambda\mu\gamma\delta}$. Similarly, the frequency with which the pairs of stimuli λ, μ such that $\frac{S_\lambda + S_\mu}{2} > \frac{S_\gamma + S_\delta}{2}$ and $\mu \; p \; \lambda$ occur will be designated $f_{\delta\gamma\mu\lambda}$.

Let us now consider those pairs of stimuli λ, μ such that $\frac{S_\lambda + S_\mu}{2} < \frac{S_\gamma + S_\delta}{2}$ but $\mu \; p \; \lambda$. In this case, this judgment may be broken up so that part of it is regarded as a judgment $\gamma \; p \; \delta$ and the other part of it as a judgment $\delta \; p \; \gamma$. Consider Fig. 3.

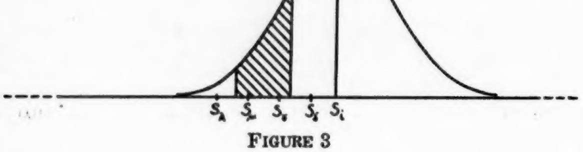


FIGURE 3

An individual who makes the judgment $\mu\ p\ \lambda$ has an attitude at the moment somewhere above $\dfrac{S_\lambda + S_\mu}{2}$. The probability that the individual would have made the judgment $\gamma\ p\ \delta$ at the time he made the judgment $\mu\ p\ \lambda$ is the cross-hatched area under the curve in Fig. 3 expressed as a fraction of the area under the curve to the right of the ordinate erected at $\dfrac{S_\lambda + S_\mu}{2}$ and is given by the equation:

$$\frac{\displaystyle\int_{X_2=\left(\frac{S_\lambda + S_\mu}{2} - S_i\right)\frac{1}{D_i}}^{X_1=\left(\frac{S_\gamma + S_\delta}{2} - S_i\right)\frac{1}{D_i}} \frac{1}{\sqrt{2\pi}} D_i\, e^{-\frac{(X-S_i)^2}{2D_i^2}}\, dX}{\displaystyle\int_{X_3=\left(\frac{S_\lambda + S_\mu}{2} - S_i\right)\frac{1}{D_i}}^{\infty} \frac{1}{\sqrt{2\pi}} D_i\, e^{-\frac{(X-S_i)^2}{2D_i^2}}\, dX}. \qquad (4)$$

The sum of the probabilities given by equation (4) for all pairs of stimuli (λ, μ) such that $\dfrac{S_\lambda + S_\mu}{2} < \dfrac{S_\gamma + S_\delta}{2}$ and $\mu\ p\ \lambda$ we shall designate $f_{\mu\lambda\gamma\delta}$.

It is immediately apparent that if the number of pairs of stimuli $\dfrac{S_\lambda + S_\mu}{2} < \dfrac{S_\gamma + S_\delta}{2}$ and $\mu\ p\ \lambda$ be designated as K, then $K - f_{\mu\lambda\gamma\delta}$ is the number of times that the judgment $\mu\ p\ \lambda$ may be taken as a judgment $\delta\ p\ \gamma$ and will be designated $f_{\mu\lambda\delta\gamma}$.

Let us now consider those pairs of stimuli $\dfrac{S_\lambda + S_\mu}{2} > \dfrac{S_\gamma + S_\delta}{2}$ but $\lambda\ p\ \mu$. In this case also, this judgment may be broken up into two parts such that part of it is regarded as a judgment $\delta\ p\ \gamma$ and the other part of it as a judgment $\gamma\ p\ \delta$.

Consider Fig. 4.

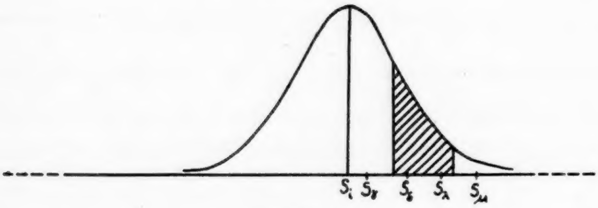


FIGURE 4

An individual who makes the judgment $\lambda\ p\ \mu$ has an attitude at the moment somewhere on the continuum below $\dfrac{S_\lambda + S_\mu}{2}$. The probability that the individual would have made the judgment $\delta\ p\ \gamma$ at the time he made the judgment $\lambda\ p\ \mu$ is the cross-hatched area under the curve in Fig. 4 expressed as a fraction of the area under the curve to the left of the ordinate erected at $\dfrac{S_\lambda + S_\mu}{2}$ and is given by the equation:

$$\frac{\displaystyle\int_{X_1=\left(\frac{S_\gamma+S_\delta}{2}-S_i\right)\frac{1}{D_i}}^{X_2=\left(\frac{S_\lambda+S_\mu}{2}-S_i\right)\frac{1}{D_i}} \frac{1}{\sqrt{2\pi}}\, D_i\, e^{-\frac{(X-S_i)^2}{2D_i^2}}\, dX}{\displaystyle\int_{-\infty}^{X_2=\left(\frac{S_\lambda+S_\mu}{2}-S_i\right)\frac{1}{D_i}} \frac{1}{\sqrt{2\pi}}\, D_i\, e^{-\frac{(X-S_i)^2}{2D_i^2}}\, dX} . \qquad (5)$$

The sum of the probabilities given by equation (5) for all pairs of stimuli (λ, μ) such that $\dfrac{S_\lambda + S_\mu}{2} > \dfrac{S_\gamma + S_\delta}{2}$ and $\lambda\ p\ \mu$ we shall designate $f_{\delta\gamma\lambda\mu}$.

Again it is immediately apparent that if the number of pairs of stimuli $\dfrac{S_\lambda + S_\mu}{2} > \dfrac{S_\gamma + S_\delta}{2}$ and $\lambda\ p\ \mu$ be designated as L, then $L - f_{\delta\gamma\lambda\mu}$

is the number of times the judgment $\lambda\ p\ \mu$ may be taken as a judgment $\gamma\ p\ \delta$ and will be designated $f_{\gamma\delta\lambda\mu}$.

The proportion of judgments $\gamma\ p\ \delta$ may now be written as a function of all the judgments as follows:

$$P_{\gamma p\delta} = \frac{f_{\gamma\delta} + f_{\lambda\mu\gamma\delta} + f_{\mu\lambda\gamma\delta} + f_{\gamma\delta\lambda\mu}}{f_{\gamma\delta} + f_{\lambda\mu\gamma\delta} + f_{\mu\lambda\gamma\delta} + f_{\gamma\delta\lambda\mu} + f_{\delta\gamma} + f_{\delta\gamma\mu\lambda} + f_{\mu\lambda\delta\gamma} + f_{\delta\gamma\lambda\mu}}, \qquad (6)$$

where $f_{\gamma\delta}$ and $f_{\delta\gamma}$ are 0 or 1 and 1 or 0, respectively, depending upon whether the judgment for the pair of stimuli γ and δ was $\delta\ p\ \gamma$ or $\gamma\ p\ \delta$.

The denominator of equation (6) is equal to the number of items. Furthermore, the proportion of judgments $\gamma\ p\ \delta$ is equal to the probability of the individual's S_i being to the left of $\frac{S_\gamma + S_\delta}{2}$, which is given by

$$P_{\gamma p\delta} = \int_{-\infty}^{X_1 = \left(\frac{S_\gamma + S_\delta}{2} - S_i\right)\frac{1}{D_i}} \frac{1}{\sqrt{2\pi}} D_i\, e^{-\frac{(X-S_i)^2}{2D_i^2}}\, dX \qquad (7)$$

and is illustrated by Fig. 5.

Hence, equation (6) may be written out in full as:

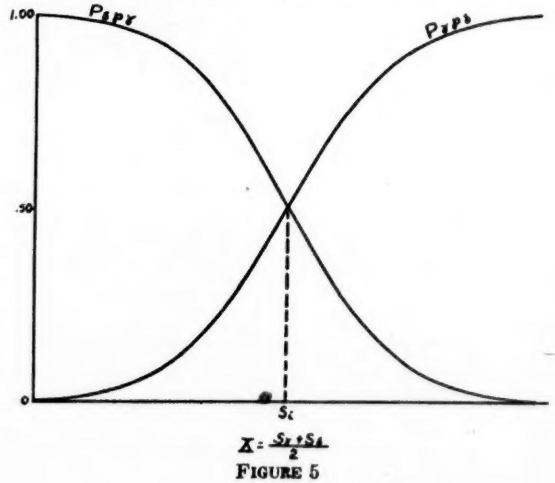


FIGURE 5

$$N \int_{-\infty}^{X_1=\left(\frac{S_\gamma + S_\delta}{2} - S_i\right)\frac{1}{D_i}} \frac{1}{\sqrt{2\pi}} D_i \, e^{-\frac{(X-S_i)^2}{2D_i^2}} \, dX =$$

$$f_{\gamma\delta} + f_{\lambda\mu\gamma\delta} + \sum_1^K \frac{\displaystyle\int_{X_3=\left(\frac{S_\lambda + S_\mu}{2} - S_i\right)\frac{1}{D_i}}^{X_1=\left(\frac{S_\gamma + S_\delta}{2} - S_i\right)\frac{1}{D_i}} \frac{1}{\sqrt{2\pi}} D_i \, e^{-\frac{(X-S_i)^2}{2D_i^2}} \, dX}{\displaystyle\int_{X_3=\left(\frac{S_\lambda + S_\mu}{2} - S_i\right)\frac{1}{D_i}}^{\infty} \frac{1}{\sqrt{2\pi}} D_i \, e^{-\frac{(X-S_i)^2}{2D_i^2}} \, dX}$$

$$+ \sum_1^L \frac{\displaystyle\int_{-\infty}^{X_1=\left(\frac{S_\gamma + S_\delta}{2} - S_i\right)\frac{1}{D_i}} \frac{1}{\sqrt{2\pi}} D_i \, e^{-\frac{(X-S_i)^2}{2D_i^2}} \, dX}{\displaystyle\int_{-\infty}^{X_2=\left(\frac{S_\lambda + S_\mu}{2} - S_i\right)\frac{1}{D_i}} \frac{1}{\sqrt{2\pi}} D_i \, e^{-\frac{(X-S_i)^2}{2D_i^2}} \, dX}, \qquad (8)$$

where N is the number of items.

Equation (8) is an implicit equation in S_i and D_i and insoluble, and only approximate solutions for S_i and D_i can be obtained.

In certain types of situations, equations (1) and (2) can be used directly without the necessity of (6) as an intermediate step. Such situations would include those psychophysical experiments in which it would be possible to present the same pair of stimuli a large number of times without the subject's recognizing the *pair*. Tests of mental ability and achievement also could be constructed which would permit the use of equations (1) and (2). They would, however, have to be especially constructed for the purpose. One could, for example, prepare an arithmetic test with 1/3 of the items at each of three levels of difficulties, or 1/5 of the items at each of five levels of difficulty, etc. With a sufficiently large number of items at each level of difficulty, the proportion of items of that difficulty which can be passed by an individual can be readily computed.

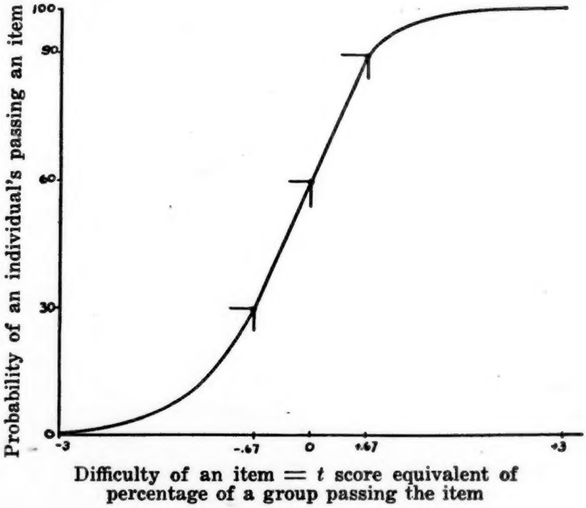


FIGURE 6

This is illustrated by a hypothetical case in Fig. 6. The three points having been obtained experimentally, there are three observation equations and two parameters, the M and σ, or the S_i and D_i, of this individual's curve.

Occasionally the use of equations (1) and (2) without equation (6) might be possible with attitude scales. If the experimenter is fortunate enough to have a lot of items end up at each of several scale values sufficiently close together, then they may be presented in all possible pairs and treated as if the same pair were being retested independently.

One clear and obvious approximation to equation (6) is to neglect the integrals. Then every pair of stimuli can be treated in turn as γ and δ and a $P_{\gamma p \delta}$ obtained for every such pair. This would give a considerable number of equations all in the same two unknowns. A solution by least squares is readily obtained.

PSYCHOMETRIKA—VOL. 13, NO. 2
JUNE, 1948

A BINOMIAL METHOD FOR ANALYZING PSYCHOLOGICAL FUNCTIONS

J. A. GENGERELLI
UNIVERSITY OF CALIFORNIA, LOS ANGELES

On the basis of the assumption that distributions of scores on psychological functions are generated by a finite number of equally probable factors, a method is presented which yields the number of factors and their probability p. The statistics β_1 and β_2 are used for this purpose. An experiment utilizing a code-transcription test is described in which the method was employed to analyze performance at several stages in the learning process. n was found to be 10 for the first 2 minutes of practice and 19 for the second 2 minutes. For the third, fourth, and fifth 2-minute periods, no value could be obtained owing to the pronounced leptokurtosis of the distributions. The first 3 periods of practice, when lumped together, gave an n of 33. It is suggested that the method offers a means of comparing the variability and "complexity" of otherwise non-comparable psychological functions. The use of the method as an instrument of investigation in the field of factor analysis is described.

1.

The distributions of scores resulting from the measurement of N subjects on a number of distinct psychological functions has always offered a challenge to the theoretical psychologist. Thus, r distinct tests can yield r means and r standard deviations, but since these measures are expressed in units which are not comparable, a rather unsatisfactory state of affairs arises from the theoretical point of view. This is especially clear in the case of standard deviations: how can it be determined, for example, whether simple reaction time is more variable than strength of grip? At the same time, it is to be noted that, although distributions of performance scores are expressed in non-comparable units, these distributions will frequently approach normality, if N is a large number and represents a good sampling. This latter phenomenon has often been interpreted to mean that in such instances the psychological function in question is dependent upon a very large, if not an infinite, number of equally probable and independent factors. Such a suggestion follows naturally from an analytical consideration of the normal probability curve. However, from the psychological point of view, it has the serious disadvantage of conferring upon all those psychological functions which are normally distributed an indeterminate degree of complexity. Thus, given two normal distributions, one of simple reaction times, and the other of scores on the Otis Intelligence Test, we should be forced to conclude that the two functions, for all practical purposes, are equally "complex."

It may be true, of course, that all normally distributed psychological functions, even the simplest, are constituted of an "infinite," or at least a very large, number of independent "factors," but this is not probable. In the present paper it will be assumed that the β_1 and β_2 of psychological functions are never in reality simultaneously equal to 0 and 3, respectively, and that in those instances where these values are obtained they are due to sampling errors. (Usually it is considered that if $\beta_1 \neq 0$ and $\beta_2 \neq 3$, the *discrepancy* is due to sampling errors.) Further, the assumption will be made that while a number of discrete independent factors may generate any given distribution of scores, this number is finite and calculable, and hence that distinct psychological functions will be generated by sets of factors which may or may not be equal in the number of their constituents. In particular, it will be assumed that any given distribution of scores, when N is large and a good sample, may be regarded as generated by a *finite* set of independent and equally probable "factors," where n represents the number of such factors, p the probability of each, and $(p+q) = 1$.

A method will be described for finding the value of p and n for any given empirical distribution of scores.

2.

It is indispensable that the determination of these two quantities should be made to depend upon statistics which are pure numbers, and hence in no way influenced by the unit of measurement used in the data. For this purpose β_1 and β_2 will be used.

In the case of the binomial, we have

$$\beta_1 = \frac{(q-p)^2}{n\,p\,q}, \tag{1a}$$

$$\beta_2 = 3 + \frac{1-6\,p\,q}{n\,p\,q}. \tag{1b}$$

Since $(p+q) = 1$, we may write (1a) and (1b), respectively, as

$$n(p-p^2)\beta_1 = (1-2\,p)^2, \tag{2a}$$

$$n(p-p^2)\,(\beta_2-3) = 1-6(p-p^2). \tag{2b}$$

We now have two simultaneous equations in p and n. Solving for p, we have

$$(\beta_2-3)(1-2p)^2-\beta_1[1-6(p-p^2)]=0\,. \tag{3}$$

Setting $(\beta_2-3)=A$, there results, after some algebraic manipulation, the following expression in p:

$$(4A-6\beta_1)p^2+(6\beta_1-4A)p-(\beta_1-A)=0\,. \tag{4}$$

Since the coefficients in (4) are the statistics β_1, β_2, and the constant 3, we can solve for p by means of the usual formula used for determining the roots of a quadratic. In this case we will have

$$p=\frac{-(6\beta_1-4A)\pm\sqrt{(6\beta_1-4A)^2+4(4A-6\beta_1)(\beta_1-A)}}{2(4A-6\beta_1)}\,. \tag{5}$$

The positive root here yields the value of q, while the negative root yields p. Having found the value of p, it is possible to substitute this value in either (1a) or (1b) (preferably both as a check against computational errors) and solve for n.

This gives us the two unknowns p and n which we seek and permits us to assert that the empirical distribution in question has values of β_1 and β_2 such that, within errors of sampling, *it may be looked upon* as generated by n independent "factors" of equal probability p. Of course, the solution of (5) may not yield a real number; the value under the radical may be negative. In this case, the assumption that the distribution may be looked upon as generated by n equally probable independent "factors" is refuted.

The procedure above is extremely susceptible to sampling errors. A sampling error affecting the value of β_1 and/or β_2 in the hundredths place may markedly alter the size of p and consequently of n; indeed, such an error may make p imaginary and hence confound the whole problem. Given the ubiquitousness of sampling errors in experimental work, however, there is not much help for this except to make N as large as possible and exercise unusual care in securing a representative sample. This great sensitivity to sampling errors is perhaps a virtue in the method rather than otherwise, since it serves to impress the worker with the importance of sampling in investigations in the field of "factor analysis,"—a fact which some workers have not fully appreciated in the past.

3.

As a first attempt in the application of the foregoing method, an experiment was performed to determine the influence of practice in a simple psychological function on the size of n. As is well known, the influence of practice on variability has attracted considerable attention in the past, but the results have always been equivocal for the lack of a suitable metric.

For this purpose a code transcription test composed of 6 elements (nonsense syllables paired with 2-digit numbers) was prepared and given to 306 college students. The key was at the top of the page and the subjects were asked to transcribe the material as rapidly as possible. The full work period of 10 minutes was subdivided into 5 periods of 2 minutes each. At the end of each 2-minute period the subjects, on command, made a mark through the last syllable written. Thus the number of items correctly transcribed during each 2-minute interval could be scored.

β_1 and β_2 were calculated for each of the 2-minute periods. In order to compare work periods of unequal length in terms of n, the scores on the first three periods were added together and β_1 and β_2 also calculated for these sums.

A second test which was considered much simpler than the preceding one was also given to the same subjects, for purposes of comparison. This involved the crossing out, by means of a diagonal line, of "0's" which filled a mimeographed page. Since this letter was the only one appearing on the page, it was considered that the test required the minimum of perceptual discrimination and involved not much more than mere speed of manual reaction. The test was of 5 minutes' duration, the subjects being instructed to work as rapidly as possible. Scores were in terms of the number of "0's" crossed out. β_1 and β_2 were calculated.*

A serious shortcoming of the experiment is the small size of the experimental group; furthermore, the group was a highly select one, namely, university students in their second and third year. As has been noted, these deficiencies may in a given case negate the whole method and, in any event, yield results with but little reliability.

4.

By substituting the various values of β_1 and β_2 in formulas (5), (1a) and (1b), the corresponding values of q and n were calculated. These are shown in the table.

It will be noted that the first and second 2-minute periods gave values of 10 and 19, respectively. The last three periods of work, how-

* S.D. in the code transcription test, for the 2-minute periods, was of the order of 8; for the cancellation test, 30.42.

	β_1	β_2	q	n
1st 2 minutes	.0182	2.82	.3956	10
2nd 2 minutes	.0433	2.94	.2916	19
3rd 2 minutes	.0072	3.09	Imaginary	—
4th 2 minutes	.0681	3.67	Imaginary	—
5th 2 minutes	.0361	3.33	Imaginary	—
1st, 2nd, and 3rd 2 minutes	.0299	2.97	.2761	33
0 — Cancellation	.0027	3.29	Imaginary	—

Since the group of subjects was not considered a reliable sample, $\sqrt{\beta_1}$ and β_2 were calculated to the nearest hundredths place only. In squaring the former, however, no digits were dropped.

The values of n are given to the nearest whole number.

ever, gave values of q which were imaginary. This latter result is associated with increased leptokurtosis in the distributions due to practice. A more satisfactory experiment would have resulted from the use of several thousand school children in one of the lower grades, or an equal number of adults within a narrow age range, chosen according to some suitable scheme of randomization. This, very probably, would have prevented a piling up of scores near the mean.

The first 3 periods when summed—that is, the scores resulting from the first 6 minutes' work—gave an n of 33. This is interesting, in that it suggests that the number of "factors" involved in 6 minutes' exercise of this function is greater than that involved in 2. Given the low reliability of our results, however, this finding is in need of corroboration.

The value of q relative to the 0-cancellation test is also imaginary. It will be noted here, also, that we have marked leptokurtosis with very little skewness in the distribution, suggesting once more a highly homogeneous group of subjects.

It may be reported, in conclusion, that the binomial method was applied to some data reported by Anastasi.* She had employed a cancellation test consisting of 27 rows of capital letters, with ten "*A's*" distributed at random throughout each row. Using over 900 college students, the subjects were asked to cross out the "*A's*," working as rapidly as possible. Two minutes were assigned to this task. This function, it would appear, is more "complex" than the one involved in our cancellation test, since perceptual discrimination is definitely required.

Anastasi reports for this test a β_1 of .0370 and a β_2 of 2.9528. These data yield a q of .2879 and an n of 24.†

*Anastasi, Anne. Practice and variability; a study in psychological method. *Psychol. Monog.*, No. 204, 1934, 45, 55 pp.

† The value of n is given to the nearest whole number.

5.

The technique described offers interesting possibilities. Not only does it afford a common metric for psychological functions for which units of measurement are non-comparable; it constitutes a useful analytical tool for the problem of the organization of mental abilities as well.

As an illustration of its use in the simplest case where but two tests are involved, x and y, let the number of elements in x be n_x, and those in y be n_y; r_{xy} is the correlation between the two tests. Now

$$r_{xy} = \frac{n_{c_{xy}}}{\sqrt{n_x \cdot n_y}}. \tag{6}$$

Since n_x, n_y, and r_{xy} are known; we may solve for $n_{c_{xy}}$, the number of elements common to the two tests.*

The values of n_i yielded by the method of binomial analysis make it possible to study the structural interrelationships obtaining between any number of tests. Consider the case of three tests x, y, and z. The application of formula (6) to n_x, n_y, and n_z yields the values $n_{c_{xy}}$, $n_{c_{xz}}$, $n_{c_{yz}}$. We require still another quantity, $n_{c_{xyz}}$, namely, the number of factors which all three tests have in common.

This quantity may be specified in terms of a maximum and a minimum value; that is, given the values n_x, n_y, n_z, $n_{c_{xy}}$, $n_{c_{xz}}$, $n_{c_{yz}}$, we may specify the maximum number of elements common to all three tests, and the minimum number, without specifying the "actual" number. In other words, given the six values indicated, there may be more than one structure or pattern which is consistent with them. This fact may be illustrated in its simplest terms by the two patterns A and B given below. Both patterns yield the same fundamental values, namely, $n_x = n_y = n_z = 3$: $n_{c_{xy}} = 1$, $n_{c_{xz}} = 2$, $n_{c_{yz}} = 1$. However, while pattern A shows one element common to all three tests, pattern B shows no element common to all three.

X		√	√	√		
Y			√		√	√
Z	√	√	√			

A

X		√	√	√		
Y		√			√	√
Z			√	√	√	

B

*In this connection an interesting check may be made on the validity of the binomial method itself. Let n_{x+y} be the number of elements resulting when the method is applied to the sums of the scores in x and y. Then n_{x+y} should equal $n_x + n_y - n_{c_{xy}}$.

The *maximum* number of elements which three tests may have in common is equal to the minimum number of elements which any two of the three tests have in common. For, in the extreme case, if there be any two tests which have no elements in common, the three tests can have none common to them all, since there are at least two tests which have none in common with one another. Similarly, if there be two tests which have but one element in common, then the three tests cannot have more than one element in common.

The determination of the *minimum* number of elements which three tests may have in common, consistent with the set of six values indicated above, is somewhat more difficult. However, this value may be obtained from the following considerations:

As an example, let $n_x = n_y = n_z = 3$; $n_{c_{xy}} = 2$, $n_{c_{xz}} = 2$, $n_{c_{yz}} = 1$. Now, if $n_{c_{xy}} = 2$ and $n_{c_{xz}} = 2$, the elements which constitute Test x have four "couplings" with the other two tests, namely, 2 with Test y and 2 with Test z. However, Test x has but 3 elements; therefore at least one of its elements must have been "coupled" twice:—once with an element in Test y and once with an element in Test z. This is equivalent to saying that there is at least one element in Test x which is common to Tests y and z. Consider next the case of Test y. This test has but 3 "couplings" with the other two tests, namely, 2 with Test x and 1 with Test z. Since this test is constituted of 3 elements, it is possible that no one of its elements is "coupled" twice; that is, it is possible that the element in y which is common to z is not among the two elements which are common to x. The same argument holds for Test z, which has 2 "couplings" with x and but one with y. However, since Test x has suffered 4 "couplings," it is evident that the system of three tests cannot have less than one element in common, given the values of the above six parameters. As a further consequence, it is evident that not all the elements in tests y and z suffered "couplings," although it was impossible to determine this from a consideration of the respective "couplings" suffered by the elements in these two tests. The conclusion can be drawn only when all the members of the "system" are considered.

Pattern C exemplifies the mutual interrelationships in this case:*

X		√	√	√	
Y			√	√	√
Z	√	√	√		

*It is interesting to note that in this instance the maximum and minimum values for the number of common elements coincide. Our analysis has shown that the number of common elements cannot be less than 1; since $n_{c_{yz}} = 1$, the number cannot be greater than 1.

The same considerations apply to situations where four or more tests are involved.

In all cases, the *maximum* number of elements common to all the tests is given by the minimum value of $n_{c_{ij}}$. The *minimum* value is obtained by calculating the ratio $\dfrac{\sum_i n_{c_{ri}}}{n_r}$ for each of the tests involved and inspecting these ratios in light of the analysis above. In this expression r designates any given test, and i denotes each of the other tests involved.

In the case of the tests w, x, y, z, we should have for Test w

$$\frac{\sum_i n_{c_{wi}}}{n_w} = \frac{n_{c_{wx}} + n_{c_{wy}} + n_{c_{wz}}}{n_w};$$

for test x

$$\frac{\sum_i n_{c_{xi}}}{n_x} = \frac{n_{c_{xw}} + n_{c_{xy}} + n_{c_{xz}}}{n_x};$$

and so on for each test in turn.

If the four tests have no elements in common, none of these ratios will exceed the critical value $K = 2$. For if any test gives a ratio greater than 2, this implies that there are more than twice as many "couplings" than elements constituting that test, and hence that at least some of its elements have more than two "couplings," thus giving the four tests some elements in common. In particular, if the ratio for Test w, for example, were 2 1/12, the numerator in the fractional remainder indicates that not less than one element in w has had more than two "couplings."

Let it be supposed, for purposes of illustration, that the ratios for tests w, x, y, and z were, respectively, 2 1/12, 2, 1 7/10, and 2 4/10. Then, w has not less than one element which has more than 2 "couplings"; x need not have any; y also may have none (indeed, it certainly has some elements which have been coupled less than twice); and z has not less than 4 elements which were "coupled" more than twice. Since one of the four tests has at least 4 elements which have been "coupled" more than twice, the four tests cannot have less than 4 common elements.

The apparent contradiction arising from the fact that different tests indicate varying numbers of elements common to the system vanishes when it is considered that the ratios are calculated and interpreted on the assumption that all the elements in any given test participate in the "couplings." This, however, may not be true in a given case. Test w, for example, might have 12 elements and a total of 25 "couplings." Assuming that each of these elements were "coupled" at least twice, then there would necessarily be one element which would be "coupled" 3 times. But, as a matter of fact, not all of w's elements may have been "coupled" twice; some may have been "coupled" but once, and others possibly not at all. The "surplus couplings" would be accounted for if 4 of w's elements were "coupled" more than twice, i.e., three times. Naturally, any one ratio cannot detect this state of affairs; it can be revealed only when the ratios of all the tests have been considered.

As has been indicated, the critical ratio K is equal to 2 in the case of 4 tests; in the case of 5 tests, $K = 3$. In general, the critical ratio is given by the expression $K = r-2$, where r represents the number of tests being considered.

In the case of 4 tests w, x, y, and z, it is possible to generate from the values n_w, n_x, n_y, and n_z supplied by the method of binomial analysis when taken in conjunction with the intercorrelations between the tests, the following set of values:—

$$\begin{array}{lll}
n_{c_{wx}} & n_{c_{wxy}} & n_{c_{wxyz}} \\
n_{c_{wy}} & n_{c_{wxz}} & \\
n_{c_{wz}} & n_{c_{wyz}} & \\
n_{c_{xy}} & n_{c_{xyz}} & \\
n_{c_{xz}} & & \\
n_{c_{yz}} & &
\end{array}$$

These values specify all the mutual relationships existing between the four tests. While the first column gives unique values, however, each of the elements in the second and third column yields two values, a maximum and a minimum.

In a similar manner tables of communalities may be derived for any number of tests. It is thus possible to place the sampling theory of mental organization on an empirical footing.

PSYCHOMETRIKA—VOL. 13, NO. 2
JUNE, 1948

AN EFFICIENT PUNCHED-CARD METHOD OF COMPUTING $\sum X$, $\sum X^2$, $\sum XY$, AND HIGHER MOMENTS

MAX E. ELLIS
OSCAR MAYER AND CO.
MADISON, WISCONSIN
AND
ARTHUR J. RIOPELLE
UNIVERSITY OF WISCONSIN

A method of computing ΣX, ΣX^2, ΣXY and higher moments on IBM equipment is described. The basic method is that of successively summary punching, collating a variable number of blank cards behind these summary cards, gang-punching the data from the summary cards into the blank cards, and totalling the entries on these summary cards. The method appears to have several advantages over those previously described, especially if coded data are used.

Introduction

In the computation of Pearson product-moment correlation coefficients and intercorrelations between several measures, much labor is expended in obtaining sums of scores, squared scores, and cross-products. The machine methods previously described have typically made use of methods of progressive digiting, master squares cards, or automatic multiplying punching.

The method described below, although it requires several types of apparatus, appears to have several advantages over those previously described. These advantages are the following:

1. The method requires no pre-punched cards.
2. Neither horizontal nor vertical digiting is required.
3. It is applicable to distributions which contain intervals of zero frequency without inserting digit cards.
4. It provides an independent internal check on the accuracy of the sums of the cross-products.
5. The values for $\sum X$, $\sum X^2$, and $\sum XY$ are obtained directly.
6. The method may be used efficiently with any number of variables.
7. Higher moments may easily be computed.

The values for $\sum X_a$, $\sum X_b$, etc., are obtained by totalizing the detail cards on the tabulator. The basic principle of computing $\sum X^2$

and $\sum XY$ is that of successively: (1) summary punching, (2) collating a variable number of blank cards behind these summary cards, (3) gang-punching the data from the summary cards into the blank cards, and (4) totalling the entries on these summary cards. The number of blank cards to be inserted behind each summary card is determined by the class interval designation.

Numerical Example

Instead of a description of the mathematical formulation of the present method, a simple numerical problem will be illustrated. Assume three variables, X_a, X_b, and X_c, where the scores on each variable range from one to five and there are ten individuals in the population. The sums of scores for each variable may be determined by simple addition. The values for $\sum X^2$ and $\sum XY$ may be obtained in the following way. Arrange the scores of variable a in descending order:

X_a	X_b	X_c
5	4	2
5	5	3
5	3	4
3	2	1
3	1	4
2	4	2
2	3	5
2	1	2
2	2	1
1	1	3

When $X_a = 5$, $\sum X_a = 15$, $\sum X_b = 12$, $\sum X_c = 9$.
" $X_a = 3$, $\sum X_a = 6$, $\sum X_b = 3$, $\sum X_c = 5$.
" $X_a = 2$, $\sum X_a = 8$, $\sum X_b = 10$, $\sum X_c = 10$.
" $X_a = 1$, $\sum X_a = 1$, $\sum X_b = 1$, $\sum X_c = 3$.

Multiplying each value of $\sum X_a$, $\sum X_b$, and $\sum X_c$ by the corresponding value of X_a, we then have:

When $X_a = 5$, $5(\sum X_a) = 75$, $5(\sum X_b) = 60$, $5(\sum X_c) = 45$.
" $X_a = 3$, $3(\sum X_a) = 18$, $3(\sum X_b) = 9$, $3(\sum X_c) = 15$.
" $X_a = 2$, $2(\sum X_a) = 16$, $2(\sum X_b) = 20$, $2(\sum X_c) = 20$.
" $X_a = 1$, $1(\sum X_a) = 1$, $1(\sum X_b) = 1$, $1(\sum X_c) = 3$.

Now summing over the whole distribution,

$$\sum_{X_a=1}^{5} (\sum X_a) = 110 = \sum X_a^2$$

$$\sum_{X_a=1}^{5}(\sum X_b)=90=\sum X_aX_b$$

$$\sum_{X_a=1}^{5}(\sum X_c)=83=\sum X_aX_c.$$

Repeating the process with respect to the values for X_b and X_c will yield similar sums of squares and cross-products for the remaining combinations.

Equipment

To perform the operations described below, the following pieces of IBM apparatus are required: sorter, type 405 alphameric electric accounting machine (tabulator), collator with card-counting device, and a summary gang-punch.

Procedure

It is desirable, though not necessary, to code the data by class intervals to minimize the number of class intervals for each variable and to minimize the magnitude of the sums resulting therefrom. The optimal number of class intervals to be used may be determined by outside considerations, but any convenient number may be used. It should be noted, however, that the number of cards required mounts rapidly if data cards contain scores greater than 20. If this condition holds, 210 cards will be required for each variable. The class intervals are numbered consecutively from one to the highest actual number of class intervals. Class interval designations of zero cannot be used. The raw scores are then transformed into interval-scores. The interval-scores for variable *a* are punched into field *a* (cols. 3 and 4) on the data cards, variable *b* into field *b* (cols. 5 and 6), etc. These are key-punched and verified in the usual manner. It is necessary that a complete set of scores be punched for each individual.

The data cards are then tabulated to determine the total for each field. These totals will represent $\sum X_a$, $\sum X_b$, etc., in terms of interval-scores.

The data cards are next sorted on field *a* and placed in descending order. They are then taken to a tabulator that is wired to control on field *a* and to total all fields including field *a*. The basic wiring diagram for controlling on field *a* and totalling all fields is presented in Fig. 1. This figure is designed to illustrate a problem containing 10 two-column fields, where the totals of each field for each control value of field *a* does not exceed four digits. This wiring arrangement is directly applicable to almost any problem where the interval scores range from 1 to 20 and there are no more than 700 cases in any single class interval of any of the variables unless they are skewed nega-

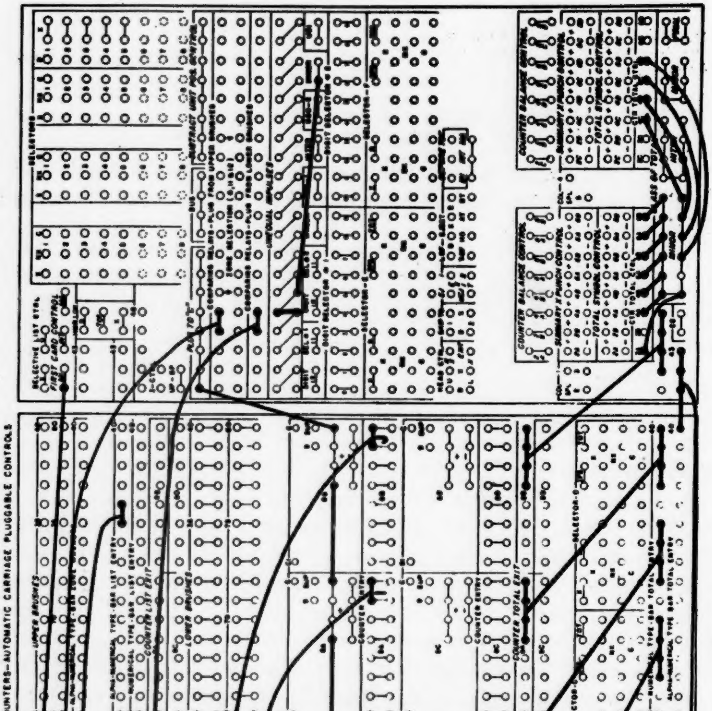


FIGURE 1
Accounting Machine Wiring Diagram

tively to an extreme degree.

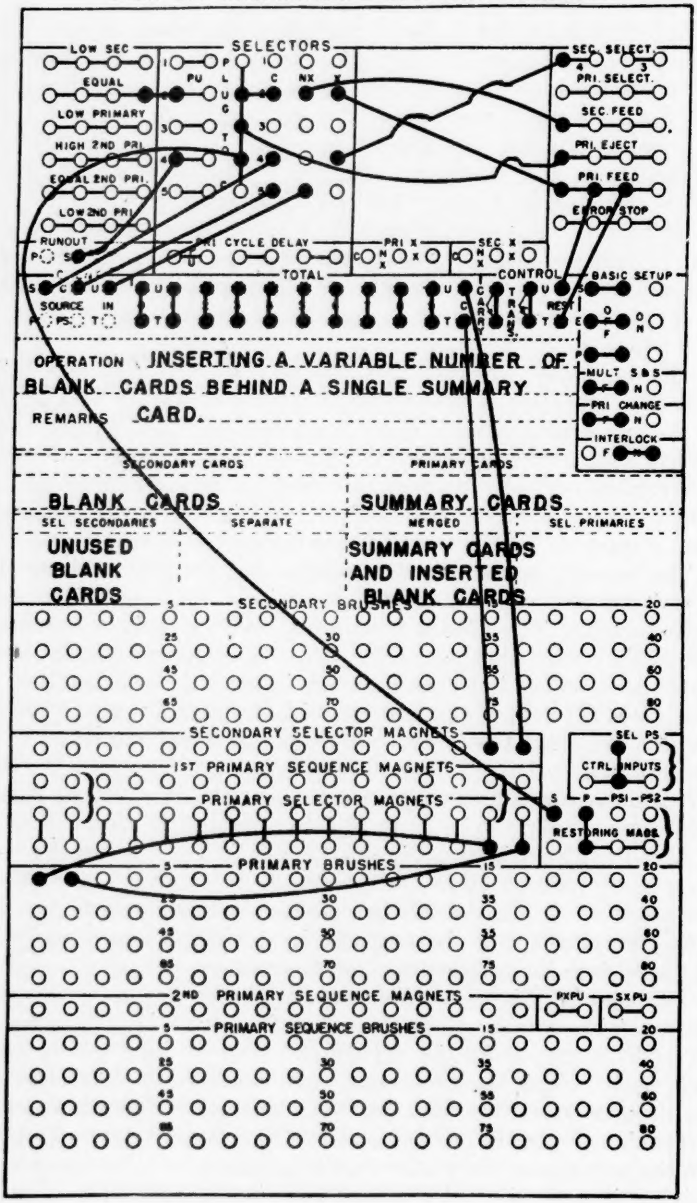
The summary punch is then connected to the tabulator and wired to punch the class-interval designation (cols. 1 and 2) and the totals of all fields for each class interval of field *a*. (In Fig. 2 the totals for field *a* are represented in cols. 3-6, field *b* in cols. 7-10, etc. The data punched into the first three fields of the sample summary card are taken from the illustrative example.) This is accomplished by minor controlling on field *a*. While summarizing, an *X*-punch (col. 78) common to all summary cards and a control field code number (cols. 79 and 80) unique to each set of summary cards are gang-punched into the summary cards. (We have used 01 as the control field code number when controlling on field *a*, 02 when controlling on field *b*, etc.)

The data cards are then re-sorted to place field *b* in descending order. The tabulator is now wired to control on field *b* by transferring the two sets of control wires from field *a* to field *b*, and a new set of summary cards is punched. Except for the gang-punching of a new control field code number into each set of summary cards, the wiring of the summary punch control panel is unchanged. This procedure is repeated for the remainder of the control fields.

The sets of summary cards are arranged in order of the control field code number and then taken to the collator equipped with a card-counting device. The collator is wired to insert a variable number of blank cards behind each summary card. The number of blank cards inserted is to be one less than the class interval designation. For example, class interval 9 would have 8 blank cards inserted. To perform this operation the collator plug-board is wired according to the plan indicated in figure 2. It will readily be seen that since the summary data are already on the summary card, the insertion of 9 cards would cause the data to be punched into a total of 10 cards. Summation over the 10 cards would in effect multiply the data on the original summary card by 10 rather than by 9. In order to make the collator insert *N*—1 blank cards behind each summary card, the deck of summary cards *must* be preceded by a single blank card before it is placed into the primary feed of the collator.

The next step is to transfer the data from each summary card to the variable number of blank cards which have been inserted behind it. This is done by the procedure known as interspersed gang-punching, by which the gang-punch will continue to punch the data from the first summary card into the blank cards following it until the machine senses the next X-punched (summary) card. The data from the second summary card will be punched into the blank cards following it until the third *X*-punched (summary) card is sensed

TYPE 077 INTERNATIONAL COLLATOR PLUGBOARD



OPERATION INSERTING A VARIABLE NUMBER OF BLANK CARDS BEHIND A SINGLE SUMMARY CARD.

REMARKS

SECONDARY CARDS	PRIMARY CARDS
BLANK CARDS	SUMMARY CARDS

SEL. SECONDARIES	SEPARATE	MERGED	SEL. PRIMARIES
UNUSED BLANK CARDS		SUMMARY CARDS AND INSERTED BLANK CARDS	

FIGURE 2
Collator Wiring Diagram to Insert N—1 Blank
Cards Behind a Single Summary Card

by the machine, etc. It is for this purpose that the X was punched into the summary cards (col. 78) while summarizing.

If the summary cards have been kept in the original order, the values for $\sum X^2$ and $\sum XY$ are obtained by totalling all fields on a tabulator set to control on the control field code number (cols. 79-80). When these totals are printed, the first line will have been controlled on field a, the second line on field b, etc. Thus a matrix of sums will result. This matrix will have as its principal diagonal the entries for $\sum X_a^2$, $\sum X_b^2$, etc. All other entries will represent sums of cross-products. It should be noted that the matrix is symmetrical about the principal diagonal. There will thus be two entries for each cross-product total. Disagreement between the two members of any pair of cross-products indicates an error.

Procedure for Computing Higher Moments

It is possible to determine the values for $\sum X_a^2 X_b$, $\sum X_a^2 X_c$, $\cdots$, $\sum X_b^2 X_a$, $\cdots$, $\sum X_a^3$, $\sum X_b^3$, etc., by a simple extension of the above procedure.

The complete set of summary cards is fed into the tabulator, which is wired to major control on the control field code number (cols. 79-80) and to minor control on the class interval designation (cols. 1-2). The data in the major and minor control fields as well as the totals for each field are summarized and punched into new summary cards. Additional blank cards are inserted behind each card in the resulting deck of summary cards in the manner previously described. Interspersed gang-punching is performed as described above. The cards are then tabulated for field totals, controlling on the control field code number (cols. 79-80).

A new matrix will result, the principal diagonal of which will contain entries for $\sum X_a^3$, $\sum X_b^3$, etc. This matrix will not be symmetrical about the principal diagonal, since the factors that are squared will correspond only to the rows and not to the columns. The value for $\sum X_a^2 X_d$ will be the fourth entry of the first row, while that for $\sum X_a X_d^2$ will be the first entry of the fourth row.

The present method is limited to the determination of the sums of products of X_a^1 by X_b^n. It does not permit the determination of $\sum X_a^n X_b^n$ when both powers are greater than one.

The tabulation of higher powers will, of course, require more cards. If the scores are not coded, the number of summary cards becomes excessive, since the number of cards used is a function of the magnitude of the scores as well as the number of variables.

The procedure is most efficient when used in the computation of higher moments.

PSYCHOMETRIKA—VOL. 13, NO. 2
JUNE, 1948

A NOTE ON THE EFFECTS OF SELECTION IN FACTOR ANALYSIS

JAMES W. DEGAN
PSYCHOMETRIC LABORATORY
UNIVERSITY OF CHICAGO

Professor Godfrey Thomson, in a communication to this laboratory, has drawn our attention to certain computational and editorial errors in a recent publication on the effects of selection in factor analysis (1). The purpose of this note is to rectify these errors. In so doing, we wish to express our gratitude to Professor Thomson for preventing their perpetuation.

Since the particular tables under consideration appear also in Chapter XIX, "The Effects of Selection," of (2) and are identical with those of (1), the material presented in this note is equally applicable to both publications. These tables were numbered identically in both (1) and (2) and may be referred to non-differentially. In order to facilitate the exposition in this note, the numbers assigned to each of the corrected tables and figure correspond to the analogous tables and figure of the original publications.

In the original article, the column vector, r_{j1}, which is used in the computation of the new communalities, following partial univariate selection, was erroneously taken from the corresponding column, c_{j1}, of the covariance matrix C_s, in Table 6. This should properly have been taken from the original correlation matrix given in Table 3. The correct substitutions have been made and the detailed computation of the new communalities is here presented in Table 8.

TABLE 8

New Communalities: $_sh_j^2 = \dfrac{h_j^2 - q^2 r_{jl}^2}{1 - q^2 r_{jl}^2}$

	h_{jl}^2	r_{jl}	r_{jl}^2	$q^2r_{jl}^2$	Num.	Denom.	$_sh_j^2$
1	.700	.837	.700	.448	.252	.552	.457
2	.700	.000	.000	.000	.700	1.000	.700
3	.700	.000	.000	.000	.700	1.000	.700
4	.700	.350	.122	.078	.622	.922	.674
5	.700	.606	.367	.235	.465	.765	.608
6	.700	.350	.122	.078	.622	.922	.674
7	.700	.606	.367	.235	.465	.765	.608
8	.700	.000	.000	.000	.700	1.000	.700
9	.700	.000	.000	.000	.700	1.000	.700
10	.700	.404	.163	.104	.596	.896	.665

Utilizing these computed values of the new communalities, the selection correlation matrix, R_s, Table 7 of the original paper, is factored. This again yields three factors and sustains Godfrey Thomson's theorem that the rank of a correlation matrix remains invariant following partial, univariate selection. This factor matrix is presented in Table 9. The principal effect of using the correct values of the new communalities is a reduction in the size of the third factor residuals. Since the centroid method of factoring was used in both instances the factor loadings are not markedly different.

Through an editorial error the values of the given factor matrix, F_0, Table 4, and the new factor matrix, F_s, Table 9, were interchanged in the original paper. The correct factor matrices, with the proper titles and table numbers are given here in Tables 4 and 9.

TABLE 4
Given Factor Matrix F_0

	I	II	III
1	.483	—.592	—.340
2	.483	.591	—.341
3	.483	.000	.683
4	.660	.216	—.467
5	.660	—.217	—.467
6	.660	—.295	.421
7	.660	—.512	.045
8	.660	.297	.420
9	.660	.512	.045
10	.837	.000	.000

TABLE 9
New Factor Matrix F_s

	I	II	III
1	.391	—.478	—.276
2	.483	.592	—.341
3	.483	.000	.683
4	.623	.303	—.440
5	.632	—.098	—.446
6	.623	—.229	.485
7	.632	—.436	.139
8	.660	.295	.421
9	.660	.512	.046
10	.808	.092	.054

Figure 4 presents the configurations of test vectors of Tables 4 and 9 illustrating the effects of selection by the method of extended vectors. It is apparent from a comparison of Figure 4 of this note and Figure 4 of the original paper that the orthogonal reference axes are in different relative positions. This change, a result of elementary orthogonal transformations, was effected for the purposes of more lucid exposition since it presents the configurations with a minimum of distortion. These transformations have resulted in numerical values in Tables 4 and 9 which are radically different from those of the original paper. It is obvious that these differences are the result of the transformations rather than sizeable computational errors.

The remainder of the original paper (1) and Chapter XIX of (2) have been checked in detail and no errors were found. It must also be emphasized that the corrections made explicit in this note in no way alter the theoretical implications or conclusions of the original paper.

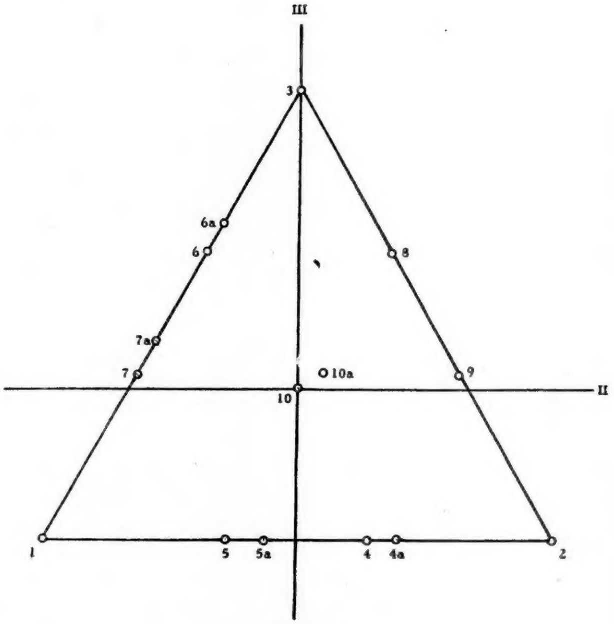


FIGURE 4

REFERENCES

1. Thurstone, L. L. The effects of selection in factor analysis. *Psychometrika*, 1945, 10, 165-198.
2. Thurstone, L. L. Multiple factor analysis. Chicago: Univ. Chicago Press, 1947, 440-472.

PSYCHOMETRIKA—VOL. 13, NO. 2
JUNE, 1948

ON THE COMPUTATION OF ZERO-ORDER CORRELATION COEFFICIENTS

CARL F. KOSSACK
PURDUE UNIVERSITY

There appears to be a gap in published computational techniques inasmuch as nowhere in the literature nor in textbooks can one find a model to be followed in computing the numerous zero-order correlation coefficients for a correlation matrix. The purpose of this paper is to present, by means of an illustration, such a model. The model consists of two computational matrices, matrix one being the Summation Matrix and matrix two the Computational Matrix. The entries on these matrices are arranged so as to facilitate the future computations.

My attention has often been drawn to the involved and unwieldy manner in which some individuals compute the zero-order correlation coefficients for a correlation matrix. An examination of statistical textbooks as well as the literature has revealed, however, no model which one might follow when first meeting this problem. Much attention is given to the problem of computing by machine methods and by other devices the necessary basic summations, and there also are many publications which show the steps necessary to transform the correlation matrix, once it has been obtained, into any of the statistical measures or groups of measures associated with it. It is the purpose of this paper to fill this gap in published computational techniques by showing, by means of an example, a method of computing all the necessary correlation coefficients at one time. It is felt that this method minimizes the time needed for such computations, at the same time enhancing the accuracy of the final results.

The computations can be accomplished by filling in but two forms. The first form I call the Summation Matrix and the second the Computational Matrix. The numerical illustration is taken from the field of Industrial Psychology* using the following variables:

X_0 = Supervisor's Rating Score obtained by using the method of paired comparisons.
X_1 = Score on Purdue Mechanical Adaptibility Test, Form A.
X_2 = Score on Adaptability Test (Tiffin and Lawshe).

* I am indebted to H. W. Porter, Purdue University, for these data.

X_3 = Objectivity Score on Guilford-Martin Personnel Inventory No. 1.

X_4 = Agreeability Score on Guilford-Martin Personnel Inventory No. 1.

X_5 = Cooperativeness Score on Guilford-Martin Personnel Inventory No. 1.

X_6 = Age at Employment.

I. The Summation Matrix

In the row and column with the common heading "Sum" is recorded the appropriate $\sum X_i$, while in the cells of the matrix are recorded the corresponding $\sum X_i X_j$. Thus, $2919 = \sum X_1$ and $18797 = \sum X_2 X_3$.

$N = 31$		X_0	X_1	X_2	X_3	X_4	X_5	X_6
	Sum	1,570	2,919	426	1,362	1,196	1,680	889
X_0	1,570	83,312						
X_1	2,919	147,978	276,787					
X_2	426	21,652	40,642	6,468				
X_3	1,362	69,938	128,913	18,797	63,950			
X_4	1,196	61,353	113,182	16,313	54,477	48,030		
X_5	1,680	86,412	158,851	23,020	76,658	67,382	98,012	
X_6	889	44,507	83,698	12,058	39,425	35,606	48,600	27,177

II. The Computational Matrix

In the column with the heading "D_i" one records the values $N\sum X_i^2 - (\sum X_i)^2$. To obtain these values one uses the edges of a plain card to mark the appropriate row and column on the Summation Matrix and simply multiplies the appropriate cell entry by N and from this product subtracts the product of the two marginal ("Sum") entries.

In the row and column with the common heading $\sqrt{D_i}$, one records the square root of the corresponding entry in the "D_i" column.

Each cell of the matrix contains three lines. On the first line of a cell one records the corresponding $N\sum X_i X_j - \sum X_i \sum X_j$, i.e., the numerator of r_{ij}.* This is obtained from the Summation Matrix by using a plain card in the same manner as one obtained the entries for the "D_i" column. On the second line one records $\sqrt{D_i} \cdot \sqrt{D_j}$, i.e., the

* Using the formula

$$r_{ij} \frac{N \Sigma X_i X_j - \Sigma X_i \Sigma X_j}{\sqrt{N \Sigma X_i^2 - (\Sigma X_i)^2} \cdot \sqrt{N \Sigma X_j^2 - (\Sigma X_j)^2}}$$

denominator, obtained from the Computational Matrix itself by again using the plain white card as a guide. On the third and last line one records the correlation coefficient obtained by dividing the entry on the second line into the entry on the first line. It is best to compute all first-line entries before starting the second line and to complete the second lines before performing the divisions.

	D_i		X_0	X_1	X_2	X_3	X_4	X_5	X_6
		$\sqrt{D_i}$	343.18	244.61	137.96	356.94	241.90	464.73	228.40
X_0	117,772	343.18							
			4,488						
X_1	59,836	244.61	83,945						
			.0535						
			2,392	16,408					
X_2	19,032	137.96	47,345	33,746					
			.0505	.4862					
			29,738	20,625	2,495				
X_3	127,406	356.94	122,495	87,311	49,243				
			.2428	.2362	.0507				
			24,223	17,518	–3,793	59,835			
X_4	58,514	241.90	83,015	59,171	33,373	86,344			
			.2918	.2961	–.1137	.6930			
			41,172	20,461	–2,060	88,238	79,562		
X_5	215,972	464.73	159,486	113,678	64,390	165,881	112,418		
			.2582	.1800	–.0320	.5319	.7077		
			–16,013	–353	–4,916	11,357	9,542	13,080	
X_6	52,166	228.40	78,382	55,869	31,510	81,525	55,250	106,144	
			–.2043	–.0063	–.1560	.1393	.1727	.1232	

PSYCHOMETRIKA—VOL. 13, NO. 2
JUNE, 1948

AN ABAC FOR THE SAMPLE RANGE

MARCO P. SCHÜTZENBERGER
PARIS, FRANCE

An abac is computed which gives the probability that at least $q\%$ of the whole population will be included in the range of a random sample of given size. Applications are suggested for testing homogeneity of a sampling from a given population.

Let a sample of N values be randomly drawn from an infinite continuous distribution. The present chart gives the probability p that at least q per cent of the whole population lie between the extreme values of the sample.

For instance, if the universe is that of the speeds in a given performance test, and the sample is constituted of N subjects passing in a given day, q will be the proportion to all subjects of those who will complete the task slower than the best of the sample and faster than the poorest.

It is easy to prove that the probability p, the percentage q, and the sample size N are related by

$$p = 1 - \left(\frac{m-1}{m}\right)^{N-1} \left(\frac{N+m-1}{m}\right)^{*}, \tag{1}$$

where

$$m = \frac{1}{1-q}. \tag{2}$$

For large enough N and m, (1) may be closely approximated by

$$p = 1 - (r+2)e^{-r-1}, \tag{3}$$

where

$$r = \frac{N}{m} - 1. \tag{4}$$

The important fact is that (1) holds for every *continuous infinite* distribution, even leptokurtic or platykurtic, skew or multimodal.

* Wilks, S. S. *Mathematical statistics.* Princeton, New Jersey: Princeton University Press, 1943. P. 93.

Use of the Abac

In ordinates are plotted, on a logarithmic scale, values of q (proportion of the whole population included in the sample range) from 50 per cent up to 998 per thousand.

In abscissas, on a logarithmic scale also, are values of the sample size N from 5 up to 1000.

Equi-probability curves are drawn for $p = 999/1000$, $99/100$, $90/100$, $75/100$, $50/100$, $25/100$, $10/100$, $1/100$, $1/1000$, so that interpolation can be made easily, and the corresponding values are written in the left and upper margins of the abac.

For example:

for $N = 20$ and $p = 99/100$, q is 75 per cent.
for $N = 200$ and $p = 999/1000$, q is 955 per thousand.

Applications

Two main applications may be suggested in the field of current psychology:

In forecasting the proportion of subjects included in the range of a given sample. For example, in a preliminary trial on 20 subjects of the speed test referred to above, extreme values of 45 and 135 seconds have been observed. Taking the equi-probability curve for $p = 99/100$, we may guess with 99 odds against one, that at least 75 per cent of all future subjects will score between 45 and 135 seconds. With 9 odds against one (that is to say, with $p = 90/100$) and 100 subjects instead of 20, this proportion would rise up to 960 per thousand.

An alternative use of the chart may be in testing a bias in sampling when there is a hint of range discrepancies and when classical tests are too difficult or lengthy to compute. For example:

Let us suppose we got from a large population the decilage of a test and that in a sample of seven, we found a subject in the first percentile and one in the last. Here q is at least 98 per cent, $N = 7$, and there is just a single chance out of a hundred ($p = 1/100$) that a random sample from the known population would be so scattered. Conversely, if in another sample of thirty subjects, we found none in the first nor in the 10th decile ($q < 80$ per cent), the probability is 99/100 that this sample would be a biased one.

Other applications of this chart could readily be found in every case where the distribution function departs from usual forms or is unknown.

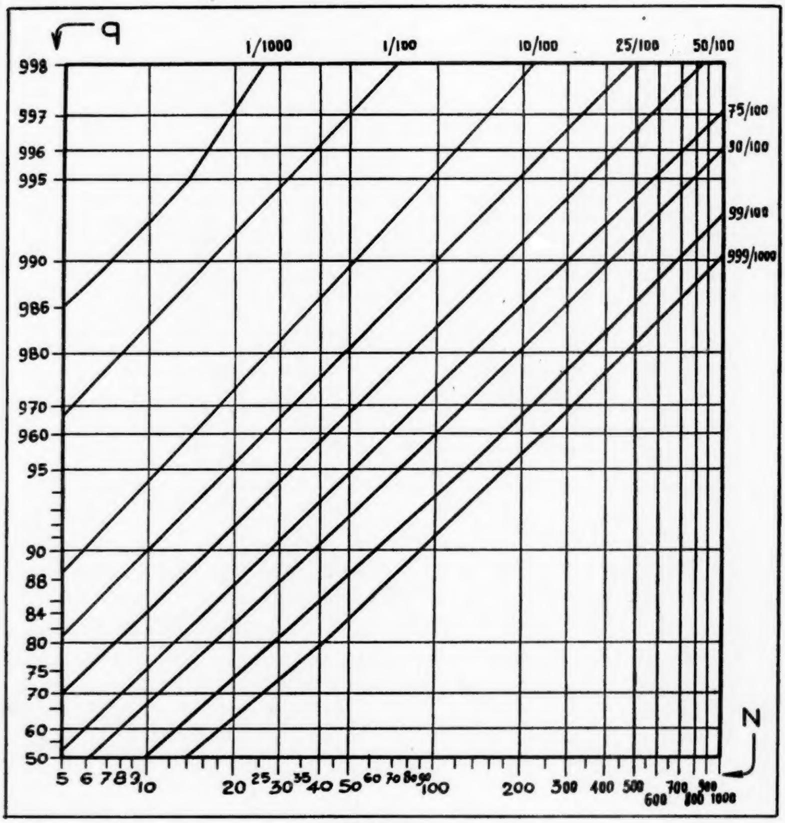

√q
1/1000
1/100
10/100
25/100
50/100
75/100
90/100
99/100
999/1000
998
997
996
995
990
986
980
970
960
95
90
88
84
80
75
70
60
50
5
6
7
8
9
10
20
25
30
35
40
50
60
70
80
90
100
200
300
400
500
600
700
800
900
1000
N

FIGURE 1

PSYCHOMETRIKA—VOL. 13, NO. 2
JUNE, 1948

SCALE ANALYSIS AND THE MEASUREMENT OF SOCIAL ATTITUDES

ALLEN L. EDWARDS AND FRANKLIN P. KILPATRICK
THE UNIVERSITY OF WASHINGTON

This paper discusses and compares the methods of attitude scale construction of Thurstone (method of equal-appearing intervals), Likert (method of summated ratings), and Guttman (method of scale analysis), with special emphasis on the latter as one of the most recent and significant contributions to the field. Despite a certain lack of methodological precision, scale analysis provides a means of evaluating the uni-dimensionality of a set of items. If the criteria for uni-dimensionality are met, the interpretation of rank-order scores is made unambiguous, and efficiency of prediction from the set of items is maximized. The Guttman technique, however, provides no satisfactory means of selecting the original set of items for scale analysis. Preliminary studies indicated that both the Likert and the Thurstone methods tend to select scalable sets of items and that their functions in this respect are complementary. A method of combining the Likert and Thurstone methods in order to yield a highly scalable set of items is outlined. Sets of 14 items selected by the method have, in the two cases where the technique has been tried, yielded very satisfactory scalability.

Three methods for the construction of attitude scales are in current use. One of these, *the method of equal-appearing intervals,* was developed by Thurstone and is described in detail in the monograph by Thurstone and Chave **(16)** which appeared in 1929. A second is the *method of summated ratings* developed by Likert and described in his monograph **(15)** published in 1932. The third is the *method of scale analysis* introduced by Guttman in an important paper **(9)** in 1944 and elaborated upon in subsequent reports **(10, 11, 12, 13, 14)**.

Method of Equal-Appearing Intervals

In the Thurstone technique approximately 100 to 200 items of opinion, bearing upon the area of content in question and ranging from extremely favorable through neutral to extremely unfavorable, are collected and edited in accordance with certain "practical criteria" (16, p. 22). These items are then presented to a judging group (usually 50 or more judges) with instructions to sort the items on a 9- or 11-point continuum. One extreme of this continuum is said to represent favorable statements, the other unfavorable statements. The middle point is said to represent a "neutral" position. The judges are requested not to react to the items in terms of their own attitudes, but merely to judge the degree of favorableness or unfavorableness of each item.

Scale values are then determined for each item by locating the point on the continuum below which and above which 50 per cent of the judges place the item. Disagreement among the judges as to the position of an item on the continuum is taken to indicate that the item is ambiguous. Ambiguity can thus be measured by the spread of the judges' ratings, and this is commonly expressed in quantitative form by means of Q, the inter-quartile range. A low Q value for an item indicates a high degree of agreement among the judges as to the degree of favorableness or unfavorableness expressed by an item. Looked at from a somewhat different point of view, if the judging group overheard an individual expressing this opinion, they would agree among themselves in locating the individual on the 9- or 11-point continuum. A statistical criterion for evaluating the "relevance" of an item is also available, but this criterion has seldom been used in the actual construction of scales. It is a neglected fact, but this criterion, developed by Thurstone (16), was one of the early attempts to arrive at, statistically, a uni-dimensional scale.

By selecting approximately 22 pairs of items, equally spaced along the continuum (insofar as this is possible) and with low Q values, two comparable forms of an attitude scale are obtained. The two forms are then administered to a new group of subjects with instructions to mark the statements "agree," "disagree," or "undecided." Scores on the scales are obtained by finding the median of the scale values of the items with which the subject agrees. Reliability coefficients are computed by correlating the scores on the two forms of the scale.

The Method of Summated Ratings

The Likert method also involves the collection of an initial large set of items which are then edited in accordance with a priori standards. These items are then presented to a group of subjects with instructions to mark the extent of their own agreement or disagreement with the statements. For this purpose 5 alternatives (strongly agree, agree, undecided, disagree, strongly disagree) are commonly used, although sometimes more and sometimes fewer categories of response are provided. The method of summated ratings is also adaptable to various forms of multiple responses other than agree-disagree. Weights of 1, 2, 3, 4, and 5 are assigned to the categories or steps, the direction of the weighting depending on whether a "strongly agree" or a "strongly disagree" response to a particular item is considered as representing a favorable attitude. About half the statements are so worded that a "strongly agree" response will be judged favorable and carry a 5 weight, and about half are worded so that a "strongly disagree" response will be judged favorable and carry a 5 weight. Scores are obtained for each subject by summating the weights of the individual item responses.

Criterion groups are established by taking the highest and the lowest 27 (or some other) per cent of the subjects in terms of total scores. Responses of the criterion groups are compared on the individual items. This is done by some one of the various techniques of item analysis. Differences between the means or medians of the two groups are commonly compared, but the responses could very well be analyzed by means of the point biserial correlation coefficient, the phi coefficient, or other statistics. The 20 to 25 most discriminating items, as indicated by the item analysis, are selected for inclusion in the attitude scale.

A new group of subjects is then tested and scores are obtained by summating the weights of the item responses. In the past, but one form of the attitude scale has been constructed and reliability coefficients have been based upon some form of the split-half technique, but it is not uncommon now to find two forms, consisting of 15 to 25 items each, constructed, making possible the calculation of a reliability coefficient based upon administration of the two forms.

Scale Analysis

Scale analysis, instead of starting with a large sample of items from the universe of content and then, in terms of statistical criteria, reducing this number to a smaller set constituting a scale, selects but 10 to 12 items from the universe of content and subjects these as a group to a test of scalability. This is the first point at which the Guttman technique departs from the Likert and Thurstone procedures. Since these 10 to 12 items are assumed to be a sample from the universe of content, it is a matter of some importance to know the basis on which they are selected.

Guttman tells us: "Whether or not a given item has the proper content defining the area remains a matter of intuitive judgment; perhaps the consensus of several people versed in the area could serve as a criterion" (12, p. 4). This is essentially what we do in collecting the 100 to 200 items ordinarily used in the Thurstone and Likert procedures. Let us assume that we have these 200 items available. From this number, how do we select the 10 or 12 items which Guttman would select? Apparently intuition and experience play a part in this selection also. Festinger indicates that he believes one should look for items "all of which are, to a large extent, rephrasings of the same thing" (8, p. 159). Guttman would say that the 10 or 12 items which seem to have the most "homogeneous content" should be selected (14, p. 461). We shall come back to this point later. For the present, let us assume that the items have been selected and are ready to test for scalability.

The items are submitted to 100 or more subjects who mark the extent of their agreement or disagreement with each item in the manner used in the Likert method of scale construction. Weights are then assigned to each response category for each item, using the successive integers beginning with zero. As in the Likert procedure, the highest weight is assigned to the most favorable response for each item. The total score for a subject is found by summing the weights of his responses to the individual items.

There are several alternative procedures from this point on, but since all "are virtually equivalent in the results they yield" (14, p. 458), we shall describe only one. This method, which Guttman has termed the "Cornell technique" (13), has certain advantages, including simplicity (14, pp. 458-459). The questionnaires are arranged in rank order according to the total scores. A table is then constructed with one column for each response category of each item and one row for each subject. For a 10-item questionnaire with 5 responses possible to each item and 100 subjects, this would mean a table with 50 columns and 100 rows.

Starting with the person having the highest score, the responses of each subject to each item are recorded by placing a check mark in the appropriate cell of the table. When completed, the table affords a record of all the available data. It is Guttman's contention that in order to call these 10 items a scale, certain conditions must be met with respect to the pattern of check marks, the most important of which is that "from a person's rank alone we can reproduce his response to each of the items in a simple fashion" (13, pp. 4-5). The other conditions which must be met will be discussed later.

Let us see what this would mean in the case of *perfect* reproducibility. Let us suppose that for the first question we have 15 individuals with weights of 4, 20 with weights of 3, 30 with weights of 2, 20 with weights of 1, and 15 with weights of 0. Now if the 4 response is judged more favorable than the 3, the 3 more favorable than the 2, and so on, then the 15 subjects in the 4 category should be the 15 subjects with the highest total rank order scores; the 20 subjects making the 3 response should occupy ranks 16 through 35, and so on for the other categories. The classification of the pattern of responses to the other 9 items could be treated in a similar fashion. But since perfect reproducibility is not to be expected in practice, it becomes a matter of interest to measure the *degree of reproducibility* present for any given set of responses.

This is accomplished by establishing *cutting points* for each response category for each item. A cuting point marks that place in the rank order of subjects where the most common response shifts from one category to the next. With overlapping between responses in different categories, some choice as to the location of the cutting points is possible. Guttman believes that they should be placed so as to minimize error (**13**, p. 16).

Between cutting points all responses would fall in the same category, *if* the scale had perfect reproducibility. Consequently, responses falling outside the column or category in which they theoretically belong may be counted as errors. The errors for each category for each item are totaled and summated for all items. Let us suppose that in our example we have a total of 100 errors and a total of 1,000 responses (100 subjects x 10 responses each). We subtract our total errors from the total number of responses and express the remainder as a percentage of the total number of responses, 900/1,000 = 90 per cent. This value, which Guttman calls the *coefficient of reproducibility,* indicates the per cent accuracy with which responses to the various items can be reproduced from the rank-order scores. It can be demonstrated, however, that this interpretation is valid only if the cutting points are located in a rigorous manner and if Guttman's rule for "minimizing error" is ignored (**5**).

Guttman at first believed that a set of items should yield a coefficient of reproducibility equal to .85 before being designated as a scale (**9**). This figure was later raised to .90, for reasons discussed in a subsequent paper (**12**, p. 7). An article by Guttman and Suchman (**10**) also deals with this point.

In case the responses recorded in the first table are not sufficiently reproducible, that is, if the coefficient of reproducibility is not equal to or greater than .90, and this will usually be the case with items with as many as 5 categories of response, a second approximation table may be constructed. Where the check marks in adjacent columns or categories seem to intertwine, these categories may be combined. In this manner, the number of response categories for the individual items are reduced so as to minimize the overlapping of check marks. The combined categories are re-weighted, that is, if we have combined categories 4 and 3, and categories 2 and 1, a response of 4 *or* 3 would now be given a weight of 2, a response of 2 *or* 1 would be given a weight of 1, and the original weight of 0 would still be given to all responses in that category.

Total scores are now recomputed for each subject on the basis of the new weights and the papers are arranged in their new rank order. A new table is prepared with columns for each of the response categories. The item described above would now be represented by only 3 columns instead of the original 5. Responses of the subjects are entered in the table, cutting points are established, errors counted, and a new coefficient of reproducibility calculated. If the coefficient of reproducibility is still not satisfactory, less than .85, successive approximations may be continued until the response categories for all items have been dichotomized.

If a coefficient of reproducibility of .85 or greater is not obtained at this point, then Guttman has an interesting suggestion to offer. If it seems that one or more sub-sets of the 10 items may scale separately, then this in turn may mean that the original universe can be broken up into sub-universes which will scale. "To test the hypothesis that a scalable sub-set is part of a scalable sub-universe, *it is necessary to show that the content of this sub-universe is ascertainable by inspection,* and is distinguished by inspection from that of the rest of the universe. The practical procedure to test this hypothesis is as follows: construct new items of two types of content, one type which should belong to the apparently scalable sub-universe, and one type which should belong in the original universe but should not belong to the scalable sub-universe. If the new items designed for the apparently scalable sub-universe do scale, and scale together with the old sub-set; and if the new items designed not to be in this sub-universe do not scale with the sub-scale; then the hypothesis is sustained that a sub-universe has been defined and has been found scalable" (12, p. 4). It might be added that the investigator could also decide that some other method of scale construction, such as the Likert or Thurstone technique, would serve his purpose, for he has no assurance that he will arrive at a scale if he does carry out Guttman's suggestion.

If a coefficient of reproducibility of 85 per cent or greater is obtained at any approximation, this constitutes evidence for the scalability of the set of items. But this is not a sufficient condition, for the simple reason that *the reproducibility of any single item can never be less than the frequency present in the modal category.* For example, if we had an item with only 2 categories of response and found .9 of the 100 subjects in one of the categories, this item would have as its *minimum* reproducibility 90 per cent. Thus, it might be possible to have a set of 10 items, each with but 2 categories of response, and each with a very high modal frequency, and these items in turn would yield—would have to yield—a very high coefficient of reproducibility. This fact should always be taken into consideration when categories are combined by the method of successive approximations. It may happen that the reduction in error is simply the result of obtaining a larger modal frequency for the various items.

Obviously, in the case of items with but 2 categories of response (agree, disagree), items for which the frequencies divide .5 and .5 are valuable in keeping the coefficient of reproducibility from being spuriously high. A similar argument applies to items with more than 2 categories of response: the more evenly distributed the frequencies are in the various categories, the less the possibility of obtaining a spuriously high coefficient of reproducibility. But it should be noted that items with non-uniform frequencies are also needed in order to obtain a range of scores. With a perfect scale and all items dividing .5 and .5, only 2 scores would be possible. Although Guttman recognizes the desirability of including in the original set of items those which will yield a wide range of marginal frequencies (12, p. 8), he fails to suggest how this is to be accomplished. As will be pointed out later, techniques of item analysis seem to be called for here.

The minimum coefficient of reproducibility which it is possible to obtain with a given set of items with known frequencies in each of the categories of response can easily be determined. Simply find the proportion of responses in the modal category for each item. Sum these values and divide by the number of items, and the resulting value indicates the minimum reproducibility present in the set of items. This coefficient may be compared with that actually observed from the set of data. The limitations and implications of this coefficent are discussed in greater detail by Guttman (14, pp. 453-454).

But even this criterion is still not sufficient to determine whether or not a set of items constitutes a scale, in the sense in which Guttman uses the term (12, p. 8; 14, pp. 452-454, 456-457). Even when a coefficient of reproducibility of 90 per cent is obtained from the data, the remaining 10 per cent error may be the result of (a) random errors and/or (b) the presence of a second variable other than the one originally defined. The presence of a possible second variable is determined by an examination of the patterns of response of the subjects to determine whether "non-scale types" exist (12, p. 8).

"Scale and Non-Scale Types"

The total number of possible types (patterns of response) is a function of the number of items under consideration and the number of categories for each item. For 10 items, each with but 2 categories of response, the number of types (scale and non-scale) is 1,024. This can easily be determined from the fact that either one of two responses to the first item may be followed by either one of two responses to the second item, and this in turn may be followed by either one of two responses to the third item, and so on. We thus have 2^{10} or 1,024 possible response patterns, generating scores ranging from 0 to 10. In general, the total number of possible types is simply the product of the number of categories of the various items. By the familiar rules of permutations and combinations, we see that only one pattern of response will result in a score of 10, while there are 10 ways in which a score of 9 may be obtained, 45 patterns which will yield a score of 8, and so on. But, by definition of a scale, there should be one and only one pattern of response for each possible score. Thus with 10 questions, each with 2 categories, we would have 11 possible scale types. In general, the number of *possible scale types,* for any set of items, may be determined by summating the number of categories for each of the items, subtracting the number of questions, and adding unity. Not all possible types, scale or non-scale, may necessarily appear in the sample of individuals under observation.

We may illustrate these procedures with a hypothetical example which, in the interests of simplicity, we shall assume consists of 3 items, of which 2 have 3 categories of response and 1 has 2.

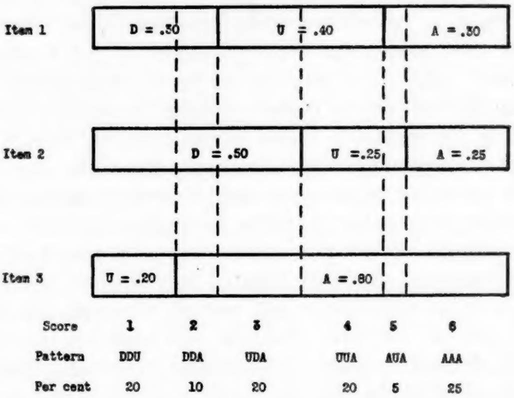
The number of possible types is $3 \times 3 \times 2 = 18$; the number of scale types would be equal to $3+3+2-3+1=6$. Let us suppose that the response categories for the first 2 items are "agree," "uncertain," and "disagree." For the third item we have the response categories "agree" and "uncertain." The weights assigned to these categories and the observed proportion in our sample making each of the various responses are as follows:

		Item 1	Item 2	Item 3
Response	Weight	p	p	p
Agree	2	.30	.25	.80
Uncertain	1	.40	.25	.20
Disagree	0	.30	.50	.00

The data above may also be arranged in the manner indicated by Fig. 1. The response patterns shown in Fig. 1 indicate the scale types and it is a simple matter to determine the relative frequency of these types for the present sample, *assuming perfect scalability.* Not all 6 scale types need appear in the sample as they do here.

Non-scale types would represent other possible patterns of response than those shown in Fig. 1, and if such non-scale types occur with substantial frequencies, then that is an indication of the presence of another variable (or other variables) (14, p. 457). The investigator might undertake to test the scalability of this second variable by constructing additional questions which seem to tap this universe of content and subjecting the set of items to a new group of respondents. This would mean, of course, that the universe of content as

FIGURE 1



Hypothetical Set of Three Items Constituting a Perfect Scale

Items 1 and 2 each have 3 categories of response: A, U, and D, with weights of 2, 1, and 0, respectively. Item 3 has only 2 categories of response: A and U, with weights of 2 and 1, respectively. The solid vertical lines mark the proportion of the sample in each category for each of the items. The possible range of scores is from 1 through 6. By the rules of permutations, 18 patterns of reponse (scale and non-scale) are possible. On the hypothesis of a perfect scale for the sample at hand, the scale patterns of response are easily determined by extending the solid vertical lines through the various items. Thus the pattern of response for "scale type 6" is AAA, for "scale type 5" it is AUA, for "scale type 4" it is UUA, and so on. The last line indicates the frequency distribution of the scale types under the conditions stated.

originally defined is not itself scalable, but that it might be broken down into sub-universes which are (14, p. 457). It is important to note that, in this case, the content of the sub-universes must be defined in such a way as to clearly indicate the separation of the two sets of items before they are treated by scale analysis (12, p. 4). In many cases, it may be feasible to ignore the presence of the second factor and to treat the original set of items as measuring a single variable.

When the patterns of response fail to indicate any substantial frequencies for non-scale types, but the coefficient of reproducibility is less than 85 per cent, the set of items is said by Guttman to constitute a "quasi-scale." Quasi-scales often show a coefficient of reproducibility that is not much higher than that predicted from the modal categories alone in the manner described previously. In quasi-scales, the error of reproducibility is assumed to be the result of a number of minor variables.

Attributes of a Scale

A set of items which meets the various criteria demanded by Guttman in order to be called a scale possesses some important properties. In the first place, the interpretation of the rank-order score is unambiguous, and it is possible to make meaningful statements about one individual being higher (more favorable) than another with respect to the variable under consideration. This would not be true of a set of items which involves more than one variable, i. e., in which one or more non-scale types are present in substantial number. Two individuals who actually differ in their attitudes on the dominant variable might obtain the same score as a result of reverse differences on the second variable. That is, one subject might stand high on the first variable and low on the second, whereas the other subject might stand low on the first variable and high on the second. Both might obtain the same score, yet the two scores are not equivalent as a measure of attitude.

Another advantage of a set of items meeting the requirements of a scale is that the zero-order correlation between an external variable and the rank-order score on the set of items will be equivalent to the multiple correlation between the individual items and the external variable. This means, of course, that efficiency in prediction from the set of items is maximized. In the case of quasi-scales, which we assume to be the result of one dominant variable and a number of minor variables, the rank order of the subjects will ordinarily be in terms of the dominant variable. If it is also true that the errors of reproducibility are random, then, regardless of the value of the coefficient of reproducibility, the zero-order correlation between an external variable and the rank-order scores will also be equivalent to the multiple correlation between the individual items and the external variable. But this equivalence will not hold in the case of sets of items where one or more non-scale types appear in substantial numbers. As pointed out earlier, this is an indication of the presence of additional variables in the set of items and the maximum efficiency of prediction could be determined only by taking into account the interrelationships between the variables present. The question as to how frequently a non-scale type must occur before a "substantial" number is reached, of course, needs to be answered precisely. Guttman's most recent publication (14) does not suggest the answer.

Selection of the Initial Set of Items

One of the troublesome problems confronting an investigator who attempts to construct a scale by following the procedures outlined only in Guttman's publications is that of selecting the initial set of items. Guttman offers little help at this point, other than to suggest that this is a matter of intuition and experience (12, p. 4). We feel that this important step should not be left a matter of intuition. On what intuitive basis, for example, did Guttman select the following 7 items (13, pp. 7-8) from the universe of content defined as attitude toward the textbook, *A Nation of Nations* (1), used in one of his classes?

1. *A Nation of Nations* does a good job of analyzing the ethnic groups in this country.

2. On the whole, *A Nation of Nations* is not as good as most college textbooks.

3. Adamic organizes and presents his material very well.

4. As a sociological treatise, Adamic's book does not rate very high.

5. Adamic does not discuss any one group in sufficient detail so that a student can obtain a real insight into problems of ethnic group relations in this country.

6. By providing a panorama of various groups, *A Nation of Nations* lets the student get a good perspective on ethnic group relations in this country.

7. *A Nation of Nations* is good enough to be kept as a textbook for this course.

This set of items was found to scale. It is conceivable, however, that at least a hundred or more items could be formulated, all of which would, in terms of a priori considerations, be judged as belonging to the universe of content as defined by Guttman. To infer, as Guttman would, that because this particular set of items scales, *any set* drawn from the universe would scale is not justifiable. We have no basis for assuming that this particular set is representative of the universe as defined. To argue that these additional items might be broken up into sets of items representative of sub-universes, and that these in turn might possibly scale, means also that the universe as originally defined (attitude toward the textbook) is not being tested with the sample set of items initially selected. The present sample would have to be regarded as a sub-universe from the original universe. And if that is so, then what is the definition of the sub-universe at hand that differentiates it from all other possible sub-universes—a step that Guttman states is essential before the sub-universe can be tested for scalability (12, p. 4)?

In many respects it is unfortunate that this problem of initial selection of items has been relatively ignored. The merits of scale analysis, as *a technique for evaluating a set of items*, are obvious and need no defense. But scale analysis can be applied to any set of items, regardless of how the set is selected. The important problem is to obtain a set which the investigator may have some assurance will meet the requirements of scale analysis. It is true, as Guttman says, "Item analysis is not adequate to test for the existence of scales in the sense of reproducibility from a single variable" (12, p. 10), but it is also true that scale analysis is not adequate for the problem of initial item selection. Guttman has not solved this problem by suggesting that we look for items with a homogeneous content (14, p. 461). Item analysis and scaling of items by the method of equal-appearing intervals have something to contribute at this point and scale analysis plays its part *after* the initial item selection.

The Scalability of Thurstone Items

Recalling that in the Thurstone technique of scale construction, items are scaled along a continuum ranging from "extremely unfavorable," through "neutral," to "extremely favorable," it is a logical conclusion that the frequency or probability of endorsement of items located along the continuum is related to the scale values, assuming a normal distribution of attitudes. This assumption has nothing to do with the *test of scalability*. The same metric that Guttman uses will apply to these items, if they prove to be scalable. It does mean, however, that for items scaled at the two extremes of the continuum, and permitting only an "agree" or "disagree" response, we would expect the modal frequencies of the items to be high and, as we move in toward the center of the continuum, we would expect the frequencies to be distributed more evenly between the two categories of response. This, if true, would provide us with some assurance that we would have a spread of marginals and also, through the inclusion of some items with a .5 and .5 division of response, a rigorous test of scalability.

There are, however, at least three difficulties with this argument as it stands. One is that if the distribution of attitudes is not normal over the entire continuum—if we have a very homogeneous group, for example, one that is strongly opposed to, let us say, capital punishment—then we would expect the majority of subjects to disagree with items that are scaled as favorable to capital punishment and the same subjects to agree with items that are scaled as unfavorable toward capital punishment. Thus our modal frequencies for all items will be quite high and our coefficient of reproducibility might not be much larger than that established as the minimum possible, i.e., not much larger than the average of the modal categories of the items.

A second difficulty is that we would expect the "neutral" items to be quite poor, in the sense of not showing clear cutting points. The reason for this, as has been demonstrated previously, is that the probability of endorsement of Thurstone "neutral" items is much the same for those with attitudes properly scaled at opposite extremes of the continuum (4). Because of the probable overlap between the responses of those with high and those with low rank-order scores on these items, responses to these items will not be reproducible from the rank-order scores. "Neutral" items, in other words, may be expected to contribute greatly to error, and the coefficient of reproducibility will be decreased accordingly.

A third difficulty is the fact that not all items with the same Thurstone scale values are equally discriminating. We have frequently found that items falling within the same scale interval and with comparable *Q* values still differ tremendously in their power to differentiate between high and low criterion groups. The hypothesis that the cutting point of an item is related to the Thurstone scale value of the item, within the limitations noted, however, is an interesting one to consider.

We had available responses of 44 subjects on the Thurstone scale measuring "Attitude Toward Capital Punishment" (17) and the responses of 55 subjects on the scale measuring "Attitude Toward Communism" (18). These papers were rescored, for purposes of scale analysis, by scoring "disagree" as 1, "?" as 2, and "agree" as 3 in the case of items scaled at the favorable end of the continuum and assigning reverse weights for items at the unfavorable end of the continuum. We used only the 12 items with scale values outside the "neutral" section of the continuum in these two tests. When the response categories were dichotomized by successive approximations, the coefficients of reproducibility were 86 per cent for the "Capital Punishment" scale and 91 per cent for the "Communism" scale.

In another test, 10 items were selected from the "Capital Punishment" scale by taking every second item in terms of scale values. Thus items ranging from the lowest scale value, through "neutral," to the highest scale value were used. The papers were rescored in the manner indicated earlier, but, as we expected, the "neutral" items (scale values of 5.3 and 5.7) failed to show any clear relationship to rank-order scores, and cutting points could not be established for them. When these two "neutral" items were eliminated and the papers were rescored and retested, the coefficient of reproducibility was 91 per cent.

On the basis of these preliminary findings, we decided to test the Thurstone items with a new and larger sample. We took the 12 items with the most extreme scale values (both high and low) from the

"Communism" scale and the 12 items with extreme scale values from the "Capital Punishment" scale. These two sets of items were given to 159 subjects. The coefficient of reproducibility for the "Capital Punishment" scale was 88 per cent and for the "Communism" scale the coefficient of reproducibility was 92 per cent. The range of the modal frequencies, however, for the "Capital Punishment" scale (all dichotomous items) was from .65 to .93 with a mean value of .82. The range for the 12 items from the "Communism" scale was from .81 to .95 with a mean value of .89. Consequently, our obtained coefficients of reproducibility do not represent any great increase over the minimum values set by the modal frequencies. The difficulty here is as pointed out earlier: our subjects were all opposed to communism and capital punishment. We need more items with lower modal frequencies.

The Scalability of Likert Items

Scale analysis can be considered as a method of item analysis applied to the multivariate distribution of a group of items. In perfect scales, for example, the 2×2 tables which can be set up to represent the interrelationships of each item with every other item would of necessity show a zero entry in one of the cells. This would mean that all of the tetrachoric *r*'s would be unity, but not necessarily all of the point coefficients. But this means that if we set up the 2×2 tables for a *large* group of items, and we then selected from this large group a *smaller* set on the basis of the values of the point coefficients, this would be an approach to the selection of a set of items which would meet the requirements of scale analysis. A similar argument could be demonstrated for other forms of item analysis, such as the relationship between response to a single item and total scores based upon all items. It would thus seem a much sounder procedure to use some form of item analysis for testing a large number of items and then to test the set of items so selected for the degree of scalability by means of scale analysis.

Since Likert items are selected on the basis of ability to differentiate between individuals with high and individuals with low total scores, it should follow that these items will tend to minimize overlap between the responses of those with high and those with low rank-order scores. And since it has already been established that Likert-selected items tend to be those falling outside the "neutral" section of the Thurstone scale continuum (4, 7), by testing a set of Likert items, we shall essentially be testing a set of Thurstone items with high and low scale values, and also with proved discriminating power.

We selected the 12 most discriminating items from a Likert scale designed to measure "Attitude Toward Labor Unions" and rescored a set of 56 papers and recorded the data in a table for scale analysis. The obtained coefficient of reproducibility was 86 per cent. The range of the modal categories for the 12 items, all with but 2 categories of response, was from .50 to .91 with a mean value of .65. Our obtained coefficient of reproducibility, 86 per cent, represents a substantial increase over the lower limit of 65 per cent.

On a second test, we rescored and recorded in a table for scale analysis the responses of 56 subjects on the 8 most discriminating items from a 21-item Likert scale designed to measure "Attitude Toward Radio." The obtained coefficient of reproducibility was 90 per cent. The range of the modal categories for the 8 items, all with but 2 categories of response, was from .68 to .95 with a mean of .81. We also selected, for purposes of comparison, the 8 least discriminating items from the 21 radio items and tested them for scalability. This time no discernible pattern appeared even after the categories of response had been reduced to dichotomies. It proved impossible to draw meaningful cutting points for these items and the coefficient of reproducibility was not computed.

A Suggested Method for Item Selection

The preliminary studies of the scalability of Likert and Thurstone items suggest a number of possible approaches to the problem of the initial selection of a set of items to be tested by scale analysis. One of the most promising with which we have been working involves first scaling a large number of items by the method of equal-appearing intervals. On the basis of scale and *Q* values, we reduce the initial set of items to a smaller number. These items are then given to another sample and subjected to item analysis. We then plot the discriminatory power of the items against the Thurstone scale values and select from within each scale interval the two or three items with the greatest discriminatory power. No items are selected from the middle or "neutral" sections of the scale continuum.

The method of item selection proposed overcomes one of the major objections to Guttman's scale analysis. It places the initial selection of the set of items to be tested for scalability on an objective basis and removes it from the realm of a priori judgment. Sets of 14 items selected by the method have, in the two cases where we have tried the technique, yielded very satisfactory coefficients of reproducibility. This research is reported in detail in another paper **(6)**.

REFERENCES

1. Adamic, L. A nation of nations. New York: Harpers, 1945.
2. Drake, St. C., & Cayton, H. R. Black metropolis. New York: Harcourt, Brace & Co., 1945.
3. Edwards, A. L., & Kenney, K. C. A comparison of the Thurstone and Likert techniques of attitude scale construction. *J. appl. Psychol.*, 1946, 30, 72-83.
4. Edwards, A. L. A critique of "neutral" items in attitude scales constructed by the method of equal appearing intervals. *Psychol. Rev.*, 1946, 53, 159-169.
5. Edwards, A. L. On Guttman's scale analysis. *Educ. psychol. Meas.* In press.
6. Edwards, A. L., & Kilpatrick, F. P. The scale-discrimination method for measuring social attitudes. *J. appl. Psychol.* In press.
7. Ferguson, L. W. A study of the Likert technique of attitude scale construction. *J. soc. Psychol.*, 1941, 13, 51-57.
8. Festinger, L. The treatment of qualitative data by "scale analysis." *Psychol. Bull.*, 1941, 44, 149-161.
9. Guttman, L. A basis for scaling qualitative data. *Amer. sociol. Rev.*, 1944, 9, 139-150.
10. Guttman, L., & Suchman, E. A. Intensity and a zero point for attitude analysis. *Amer. sociol. Rev.*, 1947, 12, 57-67.
11. Guttman, L. The desire of enlisted men for post-war full-time schooling: an example of a scale. Research Branch, Information and Education Division, Army Service Forces, Survey 63-E.
12. Guttman, L. Questions and answers about scale analysis. Research Branch, Information and Education Division, Army Service Forces, Report D-2, 1945.
13. Guttman, L. The Cornell technique for scale and intensity analysis. Mimeographed, 1946.
14. Guttman, L. On Festinger's evaluation of scale analysis. *Psychol. Bull.*, 1947, 44, 451-465.
15. Likert, R. Technique for the measurement of attitudes. *Arch. Psychol.*, 1932, No. 140, 55 pp.
16. Thurstone, L. L. & Chave, E. J. The measurement of attitude. Chicago: Univ. Chicago Press, 1929.
17. Thurstone, L. L. & Peterson, R. C. *Scale for the measurement of attitude toward capital punishment*: *Form A*. Chicago: Univ. Chicago Press, 1931.
18. Thurstone, L. L. *Scale for the measurement of attitude toward communism*: *Form B*. Chicago: Univ. Chicago Press, 1931.

PSYCHOMETRIKA—VOL. 13, NO. 2
JUNE, 1948

BOOK REVIEWS

LOUIS L. THURSTONE. *Multiple Factor Analysis.* Chicago: University of Chicago Press, 1947. Pp. 535.

A full statement of the progress made in the Chicago laboratory since the publication of Professor Thurstone's veritably classical *Vectors of Mind* has been awaited by all factor analysts with great interest. That interest has been stimulated but never entirely satisfied by the stream of publications in article form; for these articles necessarily presented methods in a state of flux and without that mutual comparison which would give a definitive statement of Professor Thurstone's final evaluation.

To say that the purpose of the present volume is to bring the *Vectors of Mind* up to date would, however, indicate only one aspect of what has been achieved. That momentous work was a pioneer study, in which Professor Thurstone voyaged on strange seas of thought very much alone. His account of what he saw was not easy to follow, for pioneers do not generally approach new ideas from the angles at which they are best presented to the student. An unexpected excellence, therefore, in the present work is the ease and fitness of its order and style of presentation from the standpoint of education. In the first place the book opens with a mathematical introduction, particularly in matrix algebra, which enables the psychology student to find his way about, in those elegant, but difficult, modes of mathematical presentation with far more profit than if he were left to roam through mathematical treatises on his own without guidance as to the relevance of theorems to the factor problem. Secondly, each statement of a general theorem, anywhere in the book, is usually accompanied by a practical and particular numerical and psychological illustration. Consequently, although this book covers the same basic mathematical notions as the *Vectors*, and some more besides, there is no excuse for any tolerably competent psychology student finding any point of insurmountable difficulty.

Following the mathematical introduction, the book begins with a chapter on "The Factor Problem" which, with a fine dignity and economy of statement, presents the basic role and purpose of factor analysis in the general framework of science, and, indeed, of philosophy. The treatment then proceeds to the fundamental factor equations, the geometric model, and the three most important and useful methods of factoring a correlation matrix. After factoring, it naturally turns to the problem of rotation—rotations for simple structure—in which several methods are demonstrated.

The vexed question of estimating the unknown communalities is then taken up, and although the Chicago laboratory has evidently given much thought to the matter, the practitioner of factor analysis is still not offered any royal road to easy estimation. Thurstone's formula 15, which amounts to a minor centroid calculation, on a small cluster of positively correlating variables, is easy to work with and, as Medland has shown empirically, gives the most consistent of nine different methods of communality estimation. With how small a correlation matrix one can use this formula without having to make a second, iterative analysis remains to be seen. Fortunately the estimation of communalities is seldom a practical difficulty with the majority of factor studies, which now frequently employ thirty, forty or more variables. The theoretical issues are, however, interesting, and Professor Thurstone takes issue with Godfrey Thomson's notion that the aim in estimating communalities is to minimize the number of common factors, asserting instead that the "problem is to determine the communalities which are implied by the given side correlations" (i.e., the matrix). He also brings Thomson's description of the puzzling Heywood case under a clearer general principle.

It is interesting—and to the present reviewer gratifying—to note that whereas Professor Thurstone stood very tentatively for oblique factors in his former volume, he has now come out wholeheartedly for following simple structure no matter to what obliqueness it leads. No one with any appreciable experience in rotating for simple structure or in seeking invariance of patterns in factors, can avoid the conclusion that the factors we find in nature are often correlated. And why should they not be, since all entities within a universe are to some extent related?

In taking to oblique factors, however, we encounter the problem of whether to apply the term "factor" to the reference vector normal to the hyperplane or to the line of intersection of the remaining hyperplanes. This does not arise among orthogonal factors, where they are identical. For many purposes—particularly in stating the immediate psychologically interesting results of a factor analysis without the labor of calculating the inverse of a matrix—the reference vector is more convenient. And although Professor Thurstone may be right that in the long run the balance of convenience lies with the alternative use of "factor," it seems to the reviewer that oblique factor methods should not be introduced without a discussion of the pros and cons of these two possible conventions.

When we cease to pay mere lip service to simple structure, and insist that it is the criterion for determining rotations, we say farewell not only to orthogonal factors but also to the ideal of wholly positive sets of factor loadings. There are still psychologists who will have difficulty over the latter. The situation seems to be that as long as research workers were concentrating on abilities it was *possible*, because of the predominantly positive correlations found there, to maintain positive loadings. It was also psychologically reasonable to do so, since, except in rare cases akin to negative transfer, it was hard to conceive of an ability being a disadvantage in *any* performance. Although Professor Thurstone has not himself worked to any extent in the personality field, he is alert to the demands which the recent applications of factor analysis to personality make upon his method. There it is reasonable to expect that personality factors will impede or suppress as well as aid performances, and Professor Thurstone has made his method flexible to this probability where many psychologists have rigidly continued to seek wholly positive loadings or strictly orthogonal factors, without regard to the natural structure. As Thurstone remarks, "The statistician must not impose the arbitrary restriction that our scientific concepts in the description of people must be uncorrelated, either in the experimental sample or in the general population."

Yet it is in regard to the simple structure principle that one may encounter the greatest disappointment in reading this book. In the first place, Professor Thurstone does not discuss any other methods, such as proportional profiles, that have been tentatively suggested for finding the given structure. This is perhaps unavoidable, since the book confines itself to the advances made at the Chicago laboratory — quite enough for one treatise — rather than to a conspectus of all develpoments in factor analysis. But since the simple structure principle — the only effective principle yet available for guiding rotation — is Professor Thurstone's own idea, one might have expected that attention would be given to developing it further and removing some of its disabilities. Its principal defect has been that occasionally two apparently equally good "simple structures" — or at least equally good hyperplane fits — can be reached from the same data.

Many people would say that its defect is the great amount of labor necessary in groping by trial and error to the required position of solution; but the labor is nothing if the ultimate solution is convincing. The task of finding a satisfactory index or expression for "goodness of simple structure" is admittedly a very difficult one, tied up as it is with the still unsolved problem of determining the distribution or probable errors of factor loadings. In default of a complete theoretical solution, however, one might perhaps have hoped justifiably for some tolerable empirical solution, based on experience, akin to Tucker's empirical formula, which has proved a useful guide in the parallel problem of determining the completeness of factor extraction. Instead, one is given a more or less arbitrary set of conditions, without adequate discussions and justification, e.g.: "Each row of the factor matrix should have at least one zero" (p. 335). Since there are bound to be a certain number of complex variables in almost any battery, this particular criterion is doubtful. At this stage of the subject, some more precise index of fit might be expected, or at least a discussion of avenues along which those with abilities and opportunities to study the matter further might profitably be searching.

Following these chapters, there is a treatment of factor invariance, which is delightfully lucid and can be read with great profit, particularly by that group of factorists still trying to obtain factorial invariance without rotation! As elsewhere, Professor Thurstone's treatment has a useful knack of dissecting out the principal source of misunderstanding and displaying it on the point of one or two well-chosen sentences. "In setting up a criterion for what is to constitute an acceptable factorial method, the chief concern is to insure that the method is adequate to discover what is determinate in the data. Factorial method has not been designed to guarantee the discovery of factors which are indeterminate in the data." It is here, in the choice of variables to give the greatest likelihood of well-determined factors, that we perceive the considerable amount of art that enters into all factor analytic research. No one who reads these chapters with understanding can continue with the misconception that factor analysis is a mechanical process of feeding variables into a mill. The treatment brings out very well the craftsmanship required in research design, the special skills that grow up in the art of factor analysis, and the interplay of psychological intuition and research design that is necessary for real advances in this field.

The new matter constituting most of the concluding part of the book concerns the intriguing developments in connection with second-order factors and the much debated issue of the effects of selection. Incidentally, Professor Thurstone sticks to the term "second-order factor" rather than "super factor," which some others have suggested. This is probably in the best interests of clarity; for there may be third- or higher-order factors also to be considered.

It would have been helpful if other examples of second-order factors, besides the box problem and the general ability factor, could have been introduced here. Our sense of the meaning of factors — and at least our ability to demonstrate that they are not such phantom ciphers as some critics suppose — would be greatly helped by a much richer accumulation of "constructed" analyses in which we know beforehand what tangible entities lie behind the facade of variables. Primary factors, we are told here, are sometimes local and incidental. On the other hand, since the correlations among the primary factors depend on relatively incidental influences of selection, one could argue also that the form if not the existence of the second-order factor is just as local and incidental. A good selection of concrete examples, not only from the physical material world but also from organic and social data, would have done much more than mathematical treatment alone to clarify, at this stage, the meaning of second-order factors.

It has long been a matter of concern to psychologists to discover how far factors would survive changes in the selection of a population and whether some might be artifacts of population sampling. In an admirably comprehensive and systematic chapter, Professor Thurstone builds on the earlier solutions of Thomson and Ledermann some gratifyingly definite conclusions. Thomson held, as Thurstone points out, a rather pessimistic view as to the maintenance of identity of factors. Thurstone shows, on the other hand, that the factor transcends changes in population sample and that if a simple structure is discoverable before selection the same structure is discoverable afterwards. Simple structure is invariant under both univariate and multivariate selection; it is only the correlations among variables and the correlations among primary factors that are modified. This supposed weakness in factorial method is strategically turned to an advantage, for "When a simple structure has been found for a test battery that has been given to an experimental population and when a plausible interpretation of the primary factors has been found, these should be regarded as hypotheses to be verified by giving the same test battery to new experimental populations that should be selected in different ways."

The critic may well ask himself whether the evident plan of this book to confine itself to the developments, methods, and approaches of the Chicago laboratory is preferable to a survey by Professor Thurstone of all kinds and conditions of contributions in this area. Interesting and provocative though the latter might be, it would not have provided a book from which one could readily teach the fundamentals of multiple factor analysis or one in which such a solid, systematic statement of the important principles could have been made. Indeed, there is something even aesthetically satisfying — in these days of formless symposia, undigested "surveys" and smug eclecticism — in the architectonic genuineness of a development that has occurred largely through the originality and unremitting toil of one man and his associates. Seen in this light, incidentally, one is not so much impressed by the number of problems that remain unsolved as by the enormous progress that has been made in the short time since Professor Thurstone's first paper on multiple factor analysis in 1931. Historians of science will be interested, in this connection, in the passing reference in the preface to the way in which the germ of this whole new development arose. "When I wrote the tetrad equation (after Spearman) . . . I discovered that the tetrad was merely the expansion of a second-order minor, and the relation (to a possible multi-factor analysis) was then obvious. If the second-order minors must vanish in order to establish a single common factor, then must the third-order minors vanish in order to establish two common factors, and so on?"

In the light of the developed treatment given in this book, that first step appears so simple — now! However, it is of more than historical interest that all of the factor analytic processes now offered for use also appear much simpler than those of a decade ago. Armed with the varied equipment of factor extraction processes, rotation methods, etc., here presented, the researcher can happily tackle matrices of variables the size and complexity of which would have appalled him in the early days of this subject. The danger points have been buoyed, the possibilities of error gauged, and shorter methods introduced without lack of accuracy.

Although the psychologist may not be aware of any paucity of psychological illustration, the fact should not be overlooked that this book is and should be written for a wider audience than psychologists alone. The whole development of this remarkable statistical tool, from the centroid method of analysis to the latest methods of rotating to simple structure, is a gift from psychology to science in general. All the newer sciences, and particularly those like sociology, meteorology, economics, and even bacteriology, which have first to discover, from a mass of uncontrollable variables, the significant entities on which to concentrate, urgently need the methods of factor analysis.

That factor analysis can proceed to set up an experiment without a particularized hypothesis seems sometimes to have puzzled the novitiate of scientific method. As Thurstone points out, "It is this latter application (discovering the nature of the underlying order) . . . that is sometimes referred to as an attempt to lift ourselves by our own boot straps . . . This is probably the characteristic of factor analysis that gives it some interest as a general scientific method." The understatement of this last sentence suggests a restrained irony at the expense of the critics of factorial method value, but the writer does not hesitate also to speak plainly on the true role of factor analysis in scientific method: "Eventually, the factorial methods, which are essentially exploratory, [are likely to] yield to the reformulation of a problem in terms of the fundamental rational constructs of the science involved. It is not unlikely that factorial analyses will point the way in the work of inventing significant and fundamental scientific concepts."

On the other hand, we are reminded, with a constant awareness of the wider possibilities, that the present factorial methods are only approximate in their formulation and assumptions. While the treatise has its feet firmly on the ground and is eminently practical in its attention to educational steps, as well as to speed and accuracy in research proceedings, it looks out, at the fringes of the known, into the middle of the unknown, for, as the author says, "It would be unfortunate if some initial success with the analytical methods . . . described here should lead us to commit ourselves to them with such force of habit as to disregard the development of entirely different constructs that may be indicated by improvements in measurement and by inconsistencies between theory and experiment."

University of Illinois RAYMOND B. CATTELL

JOHN GRAY PEATMAN. *Descriptive and Sampling Statistics.* New York: Harper and Brothers, 1947. Pp. xviii + 577.

This volume is a non-mathematical treatment of the subject matter usually included in the first-year course in psychological statistics. Computations and applications are stressed throughout. To the reviewer, this approach seems highly desirable. The result is a teachable book (exceptions noted later) for beginning students that is ample for a year's work.

The book is divided into roughly equal halves: descriptive statistics, and sampling and analytical statistics. It is debatable, however, whether this logical division is best from the pedagogical point of view. Many instructors would prefer to teach chi-square, for example, at the same time that the descriptive statistics of categorical data are discussed.

The reviewer was impressed by the following sections of the text: an initial chapter sketching the history of statistical methods; an assortment of useful statistical tables; an unusually ample treatment of categorical data and appropriate statistics; and a reasonably complete section on sampling theory and procedures. The latter features should make the book unusuallly useful to the student and instructor interested in public opinion polling. The author, as a matter of fact, draws many of his illustrations from this area of research. The illustrations, incidentally, are generally well chosen. An exception (admittedly petty) is the too frequent use of Bernreuter scores. There are some psychological tests that one would prefer to see ignored, rather than publicized, by our text-book writers.

Entirely omitted is any discussion of the analysis of variance. It was presumably considered to be beyond the scope of the volume. From the same point of view, the inadequate chapter on factor analysis might well have been omitted.

The text is marred by an unexpectedly large number of errors of a statistical rather than typographical nature. The treatment of the phi coefficient and point biserial, while not out of line with most standard texts, is inaccurate. Neither statistic positively assumes point distribution. Both should be used without corrections, even if the distributions underlying the dichotomies can be assumed to be continuous and normal, if the data have to be used in the categories correlated. Both the tetrachoric and continuous biserial correlations give an *estimate* of what the product-moment correlation would be if the data were continuously and normally distributed. Phi and point biserial *are* product-moment correlations.

This fundamental error in the treatment of these statistics leads to others. The correction applied to phi when continuity and normality can be assumed (p. 93) gives an inadequate estimate of the tetrachoric correlation and furthermore should not be applied simply because the assumptions can be made. In the formula for the standard error of the difference between proportions from correlated variables (p. 408), phi should be used in place of the tetrachoric correlation. In estimating the reliability of a test by the method of item intercorrelations (p. 476), phi coefficients should again be used rather than tetrachorics. The situations used to illustrate the continuous biserial (p. 260), prediction of a dichotomous criterion and internal consistency of test items, are actually, for most purposes, better treated by the point biserial.

Other errors noted are listed briefly below: (1) The coefficient of mean square contingency (p. 86, also p. 433) does not necessarily make the assumption of continuity. (2) Rectangular distributions (p. 115) are caused by a relatively constant level of item difficulty and relatively high item intercorrelations, neither of which can easily be associated with selective factors, the explanation advanced by the author. (3) The coefficient of relative variability frequently can not be used when the data are for the same test (p. 171), since a relatively small number of test items distorts the expected relationship between the mean and standard deviation. (4) The shape of a distribution of a sum of several parts is not determined primarily by the shape of the distributions of the parts (p. 270). The number of parts and the intercorrelations of the parts are more important. (5) Errors of measurement are not reduced by increasing the size of the sample (p. 313, also p. 326), and they do affect many statistics besides correlation coefficients, e.g., differences between means, and standard deviations. (6) Comparison of the shape of the distributions of I.Q.'s and height is hardly warranted (p. 331), since any similarity is more coincidental than fudamental. (7) The sampling distribution of t is indeed leptokurtic, but not for the reason given. The expansion of the binomial when n is small gives a distribution that is platykurtic. (p. 348). (8) The standard error presented for phi coefficient (p. 389) is incredible; the relationship to chi-square gives the reciprocal of the square root of N for a test of the null hypothesis. (9) The coefficient of relative variability is misused on page 419, and the wrong conclusion is drawn from the statistics presented. (10) The author implies (p. 474) that the relationship between length of test and reliability is restricted to the split-half procedure, and he fails to warn against the use of split-half methods for speeded tests. (11) The formula given for correcting a correlation for restriction of range (p. 480) is applicable only when the standard deviations in the restricted and unrestricted ranges are obtained for the variable not primarily restricted. (12) If an item is significantly related to total score on a test (p. 481), the test and item are necessarily reliable. (13) A partial correlation (p. 485) does not assume that any of the variables represent unitary functions or traits, and the technique is applicable and useful even though a given variable could not be held constant experimentally.

The reviewer is not an efficient proof or copy reader. For what it may be worth only one such error was discovered. On page 387 a z-value of .87 was substituted for .81, and the error was carried through the rest of the problem.

The errors listed, while unfortunate, may not entirely destroy the usefulness of the book. An instructor can sometimes impress his students by pointing out errors in the text, and the volume does contain most of the material most frequently covered in a first year course in statistics. The sections on dichotomous data and sampling, after correction of errors, would constitute valuable collateral reading in any event.

The University of Washington LLOYD G. HUMPHREYS

LEONA E. TYLER. *The Psychology of Human Differences*. D. Appleton-Century Co. Inc., 1947. Pp. xiii + 420.

The author indicates that she is writing to fulfill the needs of many classes of readers—students of psychology, individuals in related "personnel" fields of work, and "those who are interested in these problems in a non-professional way whose attitudes make up what we like to call an 'informed public opinion.'" Her stated aims are to provide up-to-date information and facts organized and presented in a form under which they can be readily assimilated. Recognizing biases and relatively unfounded opinions which prevail in this area, she has "tried wherever possible to gauge what the prevailing beliefs are and relate the research findings to these initial attitudes . . . . emphasized the methods by which dependable information can be obtained as much as the information itself . . . . attempted to synthesize and reconcile opposing points of view rather than to perpetuate old arguments . . . . to sort out the findings which stand up under critical statistical analysis from those which are in error or ambiguous, and to separate actual results from interpretations." To promote those habits of thinking about human psychological characteristics which lead to dependable information has "necessitated a thorough discussion of the basic statistical concepts, such as variability, correlation, and the significance of differences. . . . . Thus mastery of these essential skills has been brought within the intellectual range of the average college student."

The book is organized in four parts. Part I, *The Field of Differential Psychology*, includes chapters on "Developement of the basic philosophy" and "Nature and extent of measurable differences." Part II, *The Major Group Differences*, includes chapters titled "Methods and logic," "Differences between men and women," "Race and nationality differences," "Class differences," "Age differences," "The feeble-minded," and "The genius." Part III, *Factors Related to Individual Differences*, includes "Logic and statistical concepts," "The relationship of mental to physical characteristics," "The effect of practice," and "The contributions of hereditary and environmental factors to individual differences." Part IV, *The Appraisal of the Individual,* includes "The measurement of aptitudes," "The search for basic traits," and "Differential psychology—A backward and forward look."

The volume is certainly broad in scope. The author has assembled a large number of research studies and incorporated them in discussing the various problems encountered in the area of human differences. For this reason the book would seem to be best suited for a beginning course in differential psychology. There is some question, however, as to how successfully the author has "reconciled opposing points of view . . . and . . . sorted out the findings which stand up under critical statistical analysis from those which are in error or ambiguous . . ." Perhaps it is impossible to do this. The biased reader will find some results that support his opinion; the critical student will feel that he continuously comes out by the same door wherein he went. This may largely be a result of the author's attempt to present faithfully a comprehensive survey of studies made and an extremely cautious attitude in drawing conclusions, and an attempt "to separate actual results from interpretations." But few conclusions are made and most problems remain for "future research" to solve. In this reviewer's opinion, interpretations and results are not easily separated, for particular interpretations seem to depend upon particular results. Rather than separate results from interpretations it would seem more meaningful to attempt to ferret out and verbalize the many assumptions made which almost always remain implicit.

The section on "The nature of distributions" is quite good. Also commendable is the author's technique of presenting exercises at the ends of sections dealing with statistical concepts so that the reader may attempt to apply the concepts to which he has been introduced. A most creditable aim is to present the tools for evaluation in connection with the material to be evaluated. However, the author's statement that "Thus mastery of these essential skills has been brought within the intellectual range of the average college student," is open to serious question. In the discussion of the statistical significance and the standard error of the difference, the statement is made that "By the term standard error we refer to an estimate of the difference which could be expected due to chance factors alone," and then continues, "If this critical ratio is 3 or larger, it is almost certain that the groups being compared are actually different in the trait that has been measured." (p. 60). It seems doubtful, also, as to whether the discussion on correlation would result in the readers' knowing how an r is interpreted. With regard to analysis-of-variance techniques, it is stated that "They are designed to make it possible to study the effects of *several* factors simultaneously and make an analysis of their relative importance" (p. 254). Again in the discussion of percentile ranks no mention is made of an important characteristic—the inequality of scale units. (p. 336). The importance of the size of the sample is emphasized, but then there seem frequent instances where this is forgotten and tables are presented or results mentioned with no indication of the size of sample on which the data are based (e.g., pp. 291 and 374). In the discussion of factor analysis and primary mental abilities, it is stated, "McNemar has shown the existence of a very pronounced general factor running through the various problems and materials used in the 1937 Stanford-Binet test" (p. 371). McNemar, however, while utilizing the centroid method, does not rotate axes—and this "very pronounced general factor" it would seem was but the first centroid (unrotated) factor.

A rather serious limitation from this reviewer's standpoint was the absence of a definition of "intelligence." Since so much of the material related to differences in "intelligence" it would have been desirable to have this concept defined. Some of the statements made follow: "Besides showing that both men themselves and their children are differentiated according to occupational level, research has also shown specifically that intelligence-test scores made by children can be used to *predict* their later occupational level" (p. 147). "(It is to be remembered that learning ability is *not* synonymous with intelligence. Our intelligence tests measure learning ability for *only* complex, abstract sorts of material and recent work indicates that they do not predict very well the *rate* at which even this type of material will be learned)" (p. 179). "The body of research evidence which has accumulated would lead us to conclude that high intelligence, while essential to genius, is not synonymous with it" (p. 235). "There are general intelligence tests for every age and level of ability, and the nature of the trait represented by their scores has been definitely defined" (p. 391).

To sum up, that the author has succeeded in bringing together much of the information relating to human differences cannot be denied. She displays a sound, cautious attitude in interpreting most of these problems. The book will be of value to the student because it brings to his attention the great many studies that have attacked various problems in this area. While he will not become a master of the statistical techniques used to evaluate these materials, he should become cognizant of the fact that such techniques are available.

Northwestern University FRANK J. DUDEK

CLAY C. ROSS. *Measurement in Today's Schools.* Second Edition. New York: Prentice-Hall Inc., 1947. pp. xvi + 551.

This book, first published in 1941, has been revised and enlarged. Some new topics and many references have been added. A separate workbook to accompany the text is now available; consequently, tests and exercises do not appear at the end of each chapter.

The book is divided into four parts: I, the Problem of Measurement; II, the Construction of Informal Teacher-Made Tests; III, the Testing Program; IV, Measurement in Instruction. The content is designed to provide a practical introduction to the problems of educational measurement for students in teachers' colleges and schools of education. This it probably does reasonably well, though the reviewer cannot help feeling that the obvious is often belabored. To cite just one example: On page 195, where the author is discussing the administration of standarized tests, he writes, "Who should administer the tests? It goes without saying that only competent persons should administer standardized tests." The reviewer couldn't help thinking, "Then why does he say it." And this same reaction cropped up again and again throughout the book.

As stated above, the reviewer believes that the book provides a reasonably satisfactory textbook for beginning students of educational measurement, but it is regrettable that in many instances the book presents outmoded and sometimes erroneous concepts. For example, in discussing the use of the Spearman-Brown formula, the author says (p. 86), "It must be emphasized that the formula requires the use of chance halves of the test, not just any halves." Now the fact is that it is highly desirable to use halves matched for content and variability (as the author has actually suggested, in part, on the preceding page). No warning of the most severe limitation of the split-half, Spearman-Brown technique—namely, its inappropriateness for speeded tests—is given.

Statisticians will object to some of Ross's definitions and descriptions of statistical terms and procedures. The range is defined several times as "the distance between the lowest and the highest scores." (Instead of the difference between the highest and lowest scores *plus one.*) On page 248, we read, "An adequate sampling must be chosen in a random manner; . . . ." This is not always true.

In general, it must be said that this book, though originally published in 1941 and revised for republication in 1947, is of the vintage of the late thirties. No suggestion of the impact of war-time research on psychological measurement appears in the book. Consequently, the author's discussions of measurement in guidance and of other topics, such as the level of reliability required of tests used for selection purposes, seem dated.

George Peabody College for Teachers FREDERICK B. DAVIS

ALBERT B. BLANKENSHIP. *How To Conduct Consumer and Opinion Research.* New York: Harper and Brothers, 1946. Pp. xiii + 314.

This volume is intended to give a "sufficient understanding of the methods used in applying the sampling questionnaire technique to different phases of activity so that the businessman will be able to determine whether his choice of personnel for the work is a wise one." With this as his aim the author has called upon many experts in the field of market research to describe the techniques used by their organizations. The book, then, amounts to a register of the foremost survey firms, research organizations, government survey bureaus, and industrial and publishing concerns having their own survey facilities. Each author is called upon to: (1) tell about the general and specific purposes of studies in his particular field; (2) outline some of the basic methods available in survey techniques for securing the answers to the problems; and (3) cite a particular study, giving its principal points of procedure, its major results, and conclusions.

Quantitatively, the book offers little other than numerical and percentage breakdowns of results. Notable exceptions are the two chapters by the Psychological Corporation's Albert Freiberg, "Psychological Brand Barometer" and "Copy Testing." In these are quoted figures on reliability and validity of the techniques discussed. In the case of the Brand Barometer a two-day sampe-resample shows an agreement above 85% for each of the nineteen different products.

Validity was measured by comparing housewives' responses to Barometer questions with sales slips at the local store. With data from 70 homes, the 13 products checked each showed an agreement of 62% or better, with a median percentage at 72 and the highest (bread) at 100. Freiberg points out that these validity figures which show agreement between *individual* sales slips and the *corresponding* Barometer interview will be lower than the agreement between sales slips and the *entire group* responses for the reason that those "non-identical individual responses" tend to distribute themselves in relation to brand popularity.

University of Southern California CLARK L. WILSON, JR.